BITTIGER

# DS306 数据科学面试 - AB Test专题

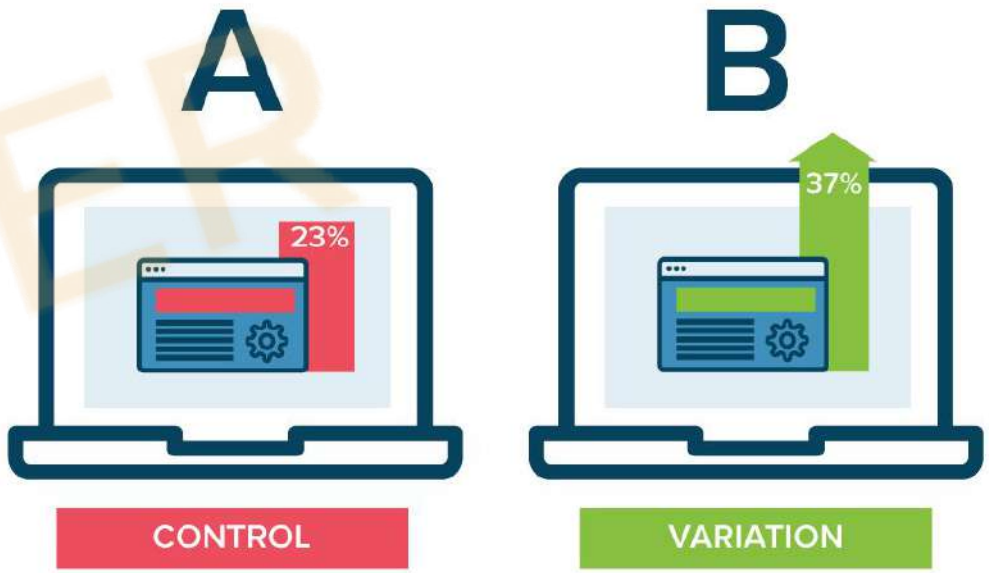
# 你的专业背景？

# Do you know how to answer the following questions?

- Survey showed teenager users are less engaged on your app after their parents join. What to do?
- You have 1M budget to spend on holiday campaigns. Possible investments include mailed ads / emails / display ads / search engine / social media ads. How would you optimize the budget?
- An engineer suggest promoting new sellers on your website to boost seller growth. But another engineer is worried about it hurting overall sales. How would you make a decision?

# Outline

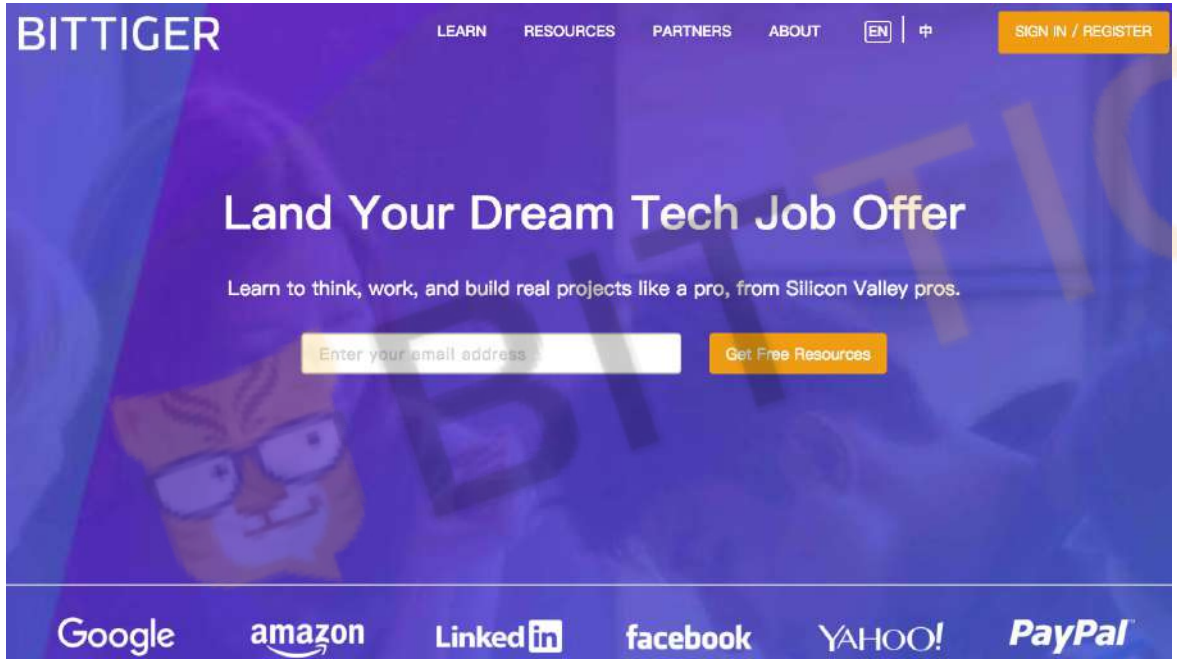
- Introduction
- Statistical Foundations
- A/B test process
  - Design of Experiment
  - Result Measurement
- Advanced A/B test topics
- A/B Test Interviews



# Introduction

What is A/B Test?

A/B testing is general framework of hypothesis testing between two groups

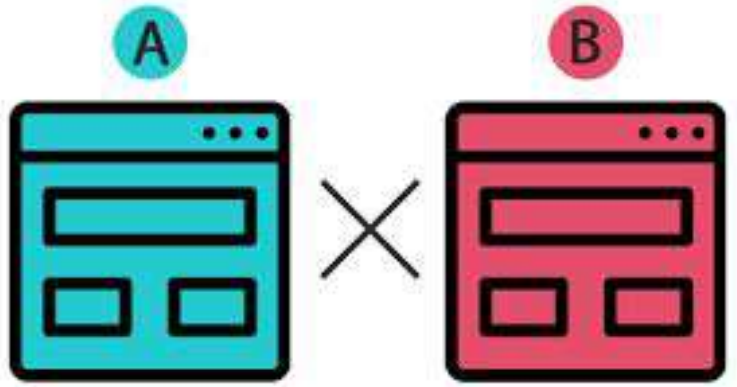



# Introduction

Why do we need A/B Test?

The goal is to

- Establish causal relationship between actions and results
- Measure impact solely from the change



# Introduction

Where is A/B Test used?

Widely used in high tech industry. Major use cases

- Product iteration

  Examples:

  - Front End: change UI design, user flow, add new features
  - Algorithm Enhancement: recommendation system, search ranking, ads display
  - Operations: define coupon value, promotion program

- Marketing optimization

  Examples

  - Search engine optimization (SEO)
  - Campaign performance measurement

In other industries, there are other forms of experiments / tests (e.g. clinical experiments in biostatistics). We will focus on A/B test in tech industry for this class

# Introduction

What's data scientist's role in A/B test?

Product Team

- Design of Experiment
- Result Measurement / Analysis
- Product Insights
- Launch Decisions


Platform Team

- Design a A/B testing platform
- Define methodology

# Statistical Foundations

# Statistical Foundations Outline

- Normal Distribution
- Central Limit Theorem
- Correlation != Causation
- Hypothesis Testing
  - T test Z test
  - P value
  - Two sample / One sample / Paired

# Normal Distribution

- Normal distribution

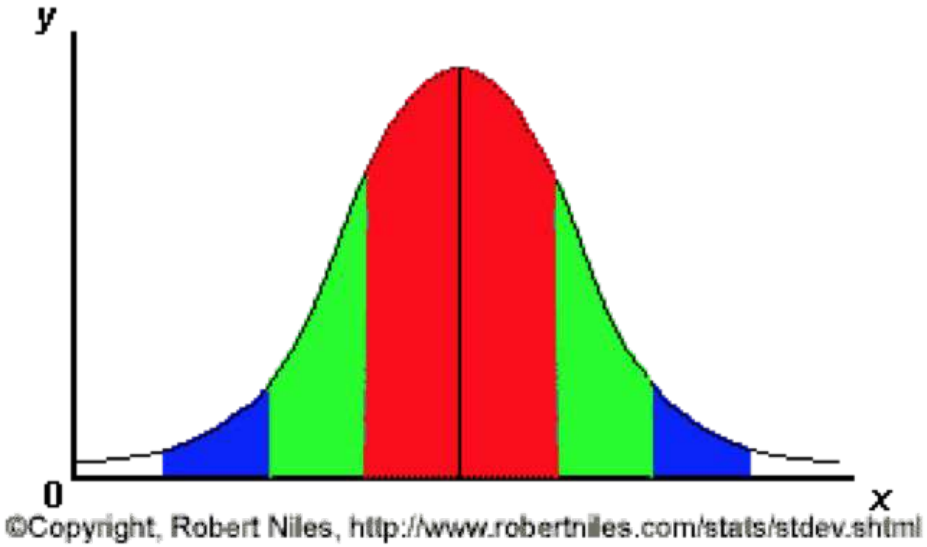
$$P(X = x) = (2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2}$$

$$E[X] = \mu \text{ and } Var(X) = \sigma^2$$

- Standard normal distribution (Z) $\mu = 0$ and $\sigma = 1$

$$X = \mu + \sigma Z \sim \mathrm{N}(\mu, \sigma^2)$$

- Beauty of normal curve (6σ)
  - [μ - 3σ, μ + 3σ] covers 99.7%
  - [μ - 2σ, μ + 2σ] covers 95%
  - [μ - σ, μ + σ] covers 68%



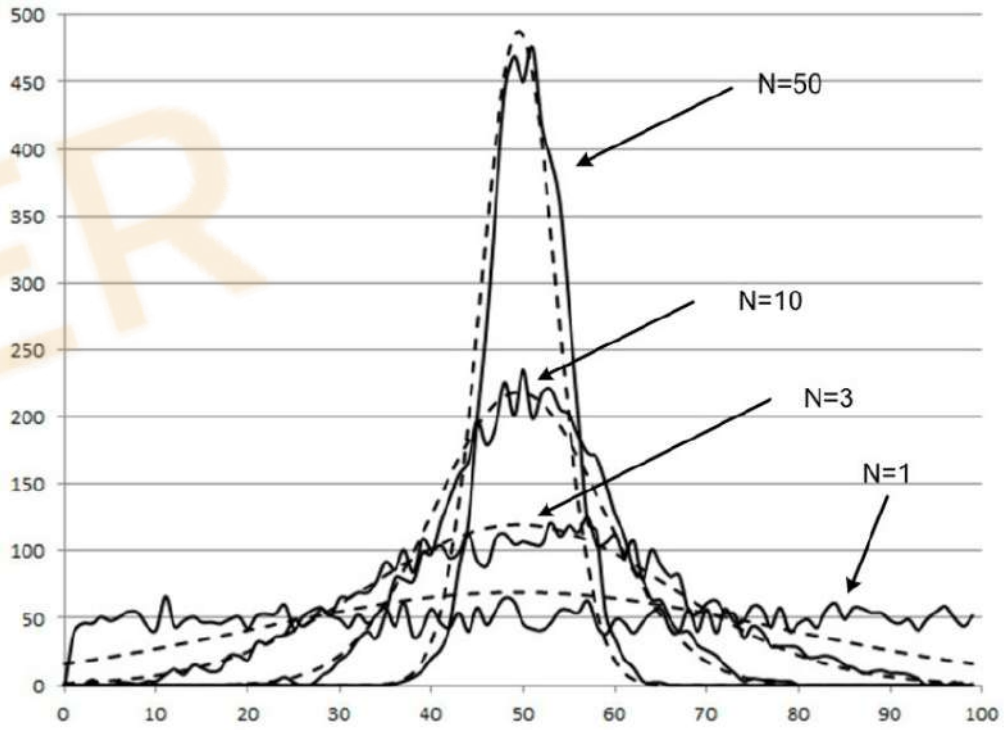
# Central limit theorem

- $X_1, X_2 \ldots X_n \ldots$ are independent, identically - distributed (IID) random variables, $X_i$ has finite mean μ and variance $\sigma^2$

$$\bar{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$$

(replacing σ by sample standard deviation, CLT still holds)

- Application
  - Binomial distribution



# CLT - Interview Quiz

User CTR of your website is p

You sampled 1000 users from your website. What is the sampling distribution of the sample mean?


$$\bar{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$$

# When CLT doesn't hold?

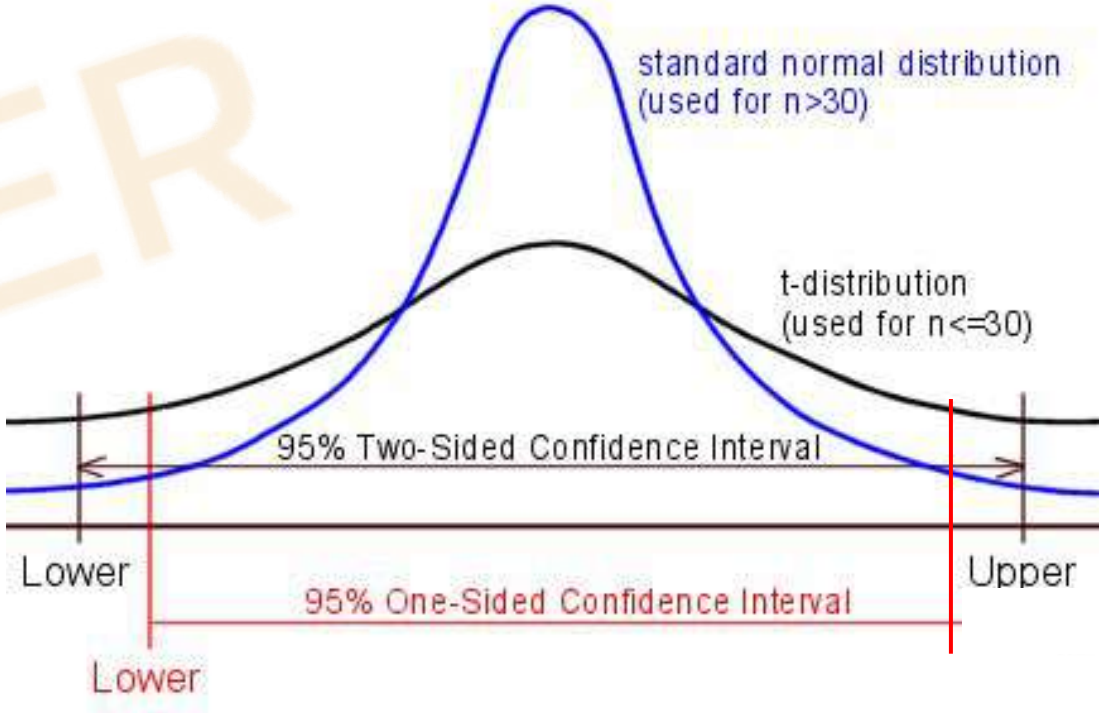
- Normal distribution -> T distribution when N < 30
  - T distribution has only one parameter: degree of freedom (df = N-1)
  - Approximate normal as df increases
  - CI under normal distribution

$$Mean_{estimate} \pm z_{1-\alpha/2} * StdErr_{estimate}$$

  - CI under t distribution

$$Mean_{estimate} \pm t_{n-1} * StdErr_{estimate}$$

  - z or t?



# Causal inference

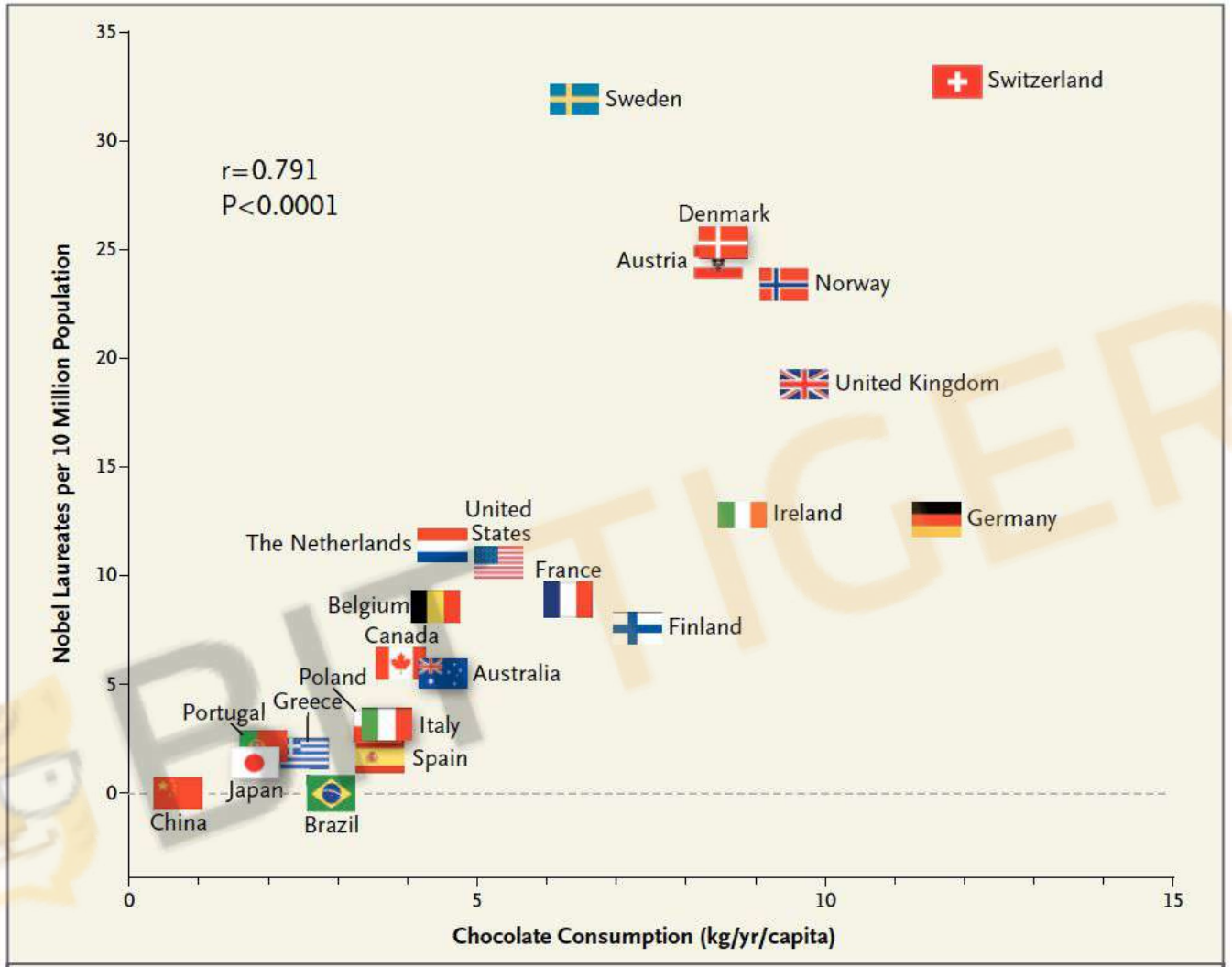



**Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

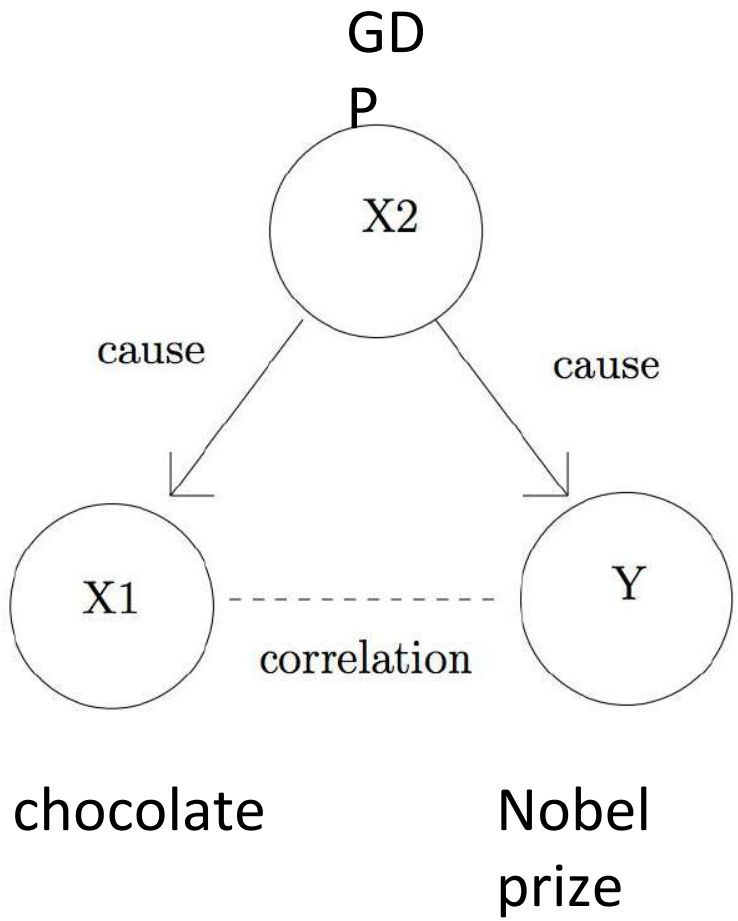
Strong correlation between chocolate consumption & Nobel laureates.

Does eating chocolate making people more likely to win Nobel Laureates?

# Correlation is not causality

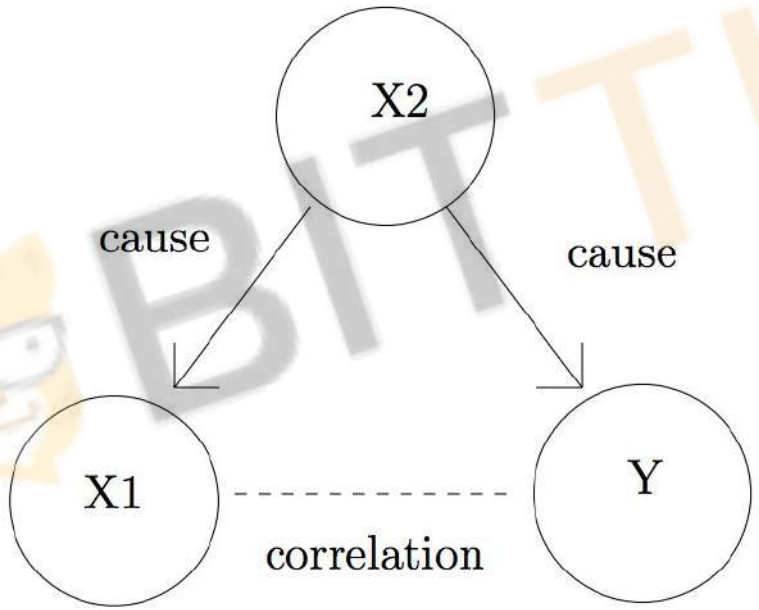
- Quote from chairman of Nobel chemistry committee
  - “Chocolate is a luxury. Wealthy individuals are more likely to be able to afford it.

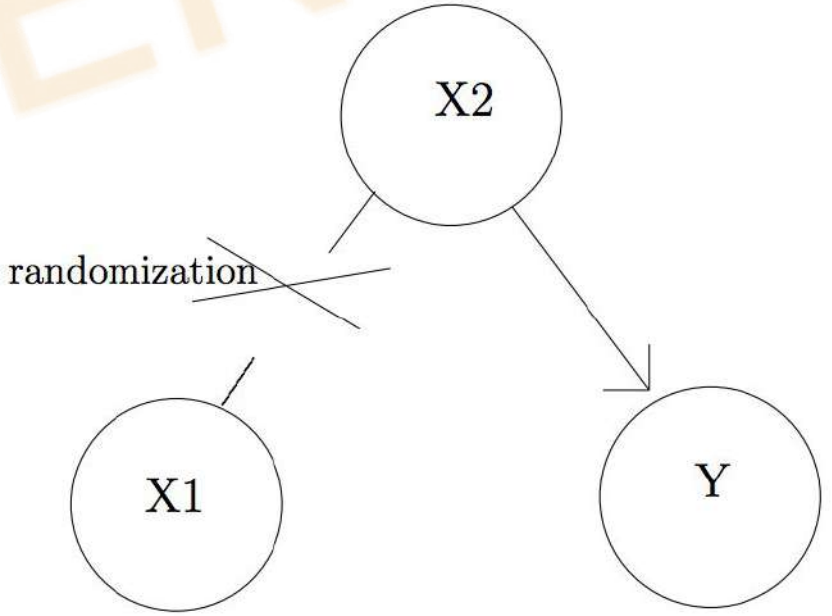
    Education is also a luxury. Poor people can’t afford to go to college for 10 years to get a PhD in chemistry. But you can’t win the Nobel prize in chemistry unless you’re a chemist. “
- Common factor
  - “GDP or wealth of a country will be correlated both with chocolate eating and with Nobel prizes. ”



# Observational study v.s. Randomized experiment

- Observational studies can suggest good experiments to run, but can't definitively show causality.
- Randomization can eliminate correlation between $X_1$ and Y due to a different cause $X_2$(confounder).



# Design randomized experiments

- Define the causal relationship to be explored, X -> Y
  - New UI decreases user interaction
- Define metric (Y)
  - Number of posts per user per day
- Design randomized experiments (A/B test)
  - Two groups of users, comparable
  - control group: old UI, experiment group: new UI
- Collect data and conduct **hypothesis testing**
  - Compare the metrics using two sample t test
- Draw conclusion

# Hypothesis testing

- Definition
  - Use sample of data to test an assumption regarding a population parameter, which could be
    - A population mean $\mu$
    - The difference in two population means, $\mu 1 - \mu 2$
    - A population variance
    - The ratio of two population variances
    - A population proportion p
    - The difference in two population proportions, $p1 - p2$

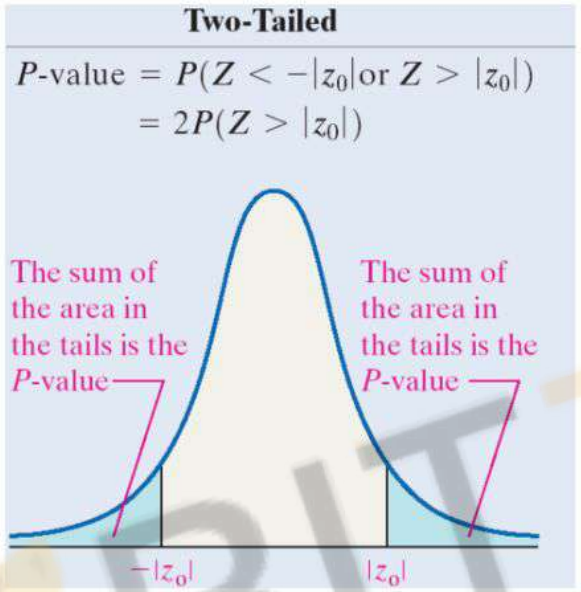
# Hypothesis testing (cont'd)

- Definition
  - Two opposing hypotheses about a population
    - Null hypothesis, $H_0$ , is usually the hypothesis that sample observations result purely from chance.
    - Alternative hypothesis, $H_1$ or $H_a$, is the hypothesis that sample observations are influenced by some non-random cause.

# Different types of alternative hypothesis

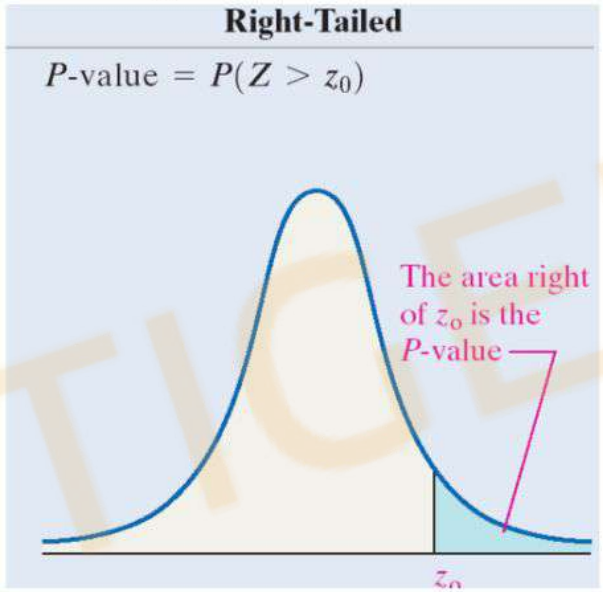
- Two tailed v.s. one tailed

Two-tail



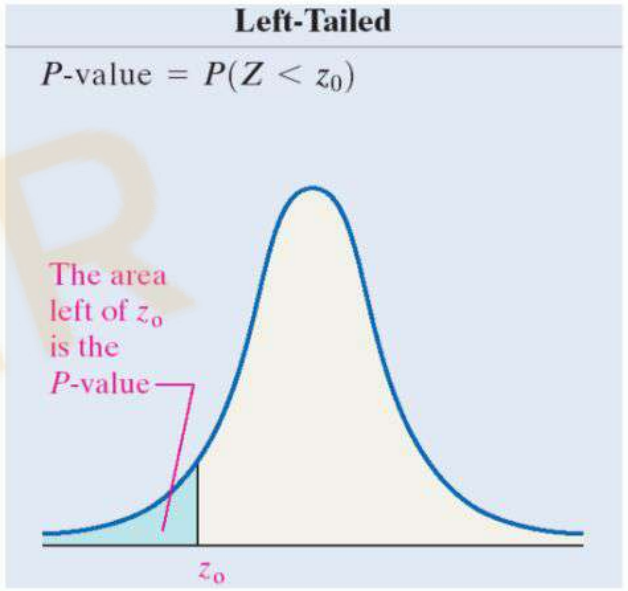

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Right Tail




$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0$$

Left Tail




$$H_0 : \mu = \mu_0$$
$$H_1 : \mu < \mu_0$$

- When to use one sided test? Reading

# How to construct hypothesis testing

Scenario: flip coin 50 times and see 40 heads. Is this a fair coin?

1 $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$

Given sample data

p is population statistic of our interest

2 $H_0 : p = \frac{1}{2}$ $H_1 : p \neq \frac{1}{2}.$

Null hypothesis

Alternative hypothesis

3 How to decide if we should reject or not reject null hypothesis?

# How to decide reject or not reject?

P_value: Assuming null hypothesis is true, what's the probability of observing a as extreme or more extreme test statistics as the observed case

Example:

flip coin 10 times, see 7 heads

P_value = Observe (>=7 heads | flip fair coin 10 times)

Flip coin 100 times, see 77 heads

P_value = Observe (>=77 heads | flip fair coin 100 times)

Two ways to decide

- "Critical value" approach, compare z with critical value
- "P value" approach, compare p value with threshold (type I error)

# Z test v.s. T test

## Hypothesis test for the population mean

- If population variance $\sigma^2$ is known and n is large, z test
- If the population variance $\sigma^2$ is unknown (most of the time), t test

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Reject H0 if z > z*

$$t = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

Reject H0 if t > t*

# One sample v.s. Two sample tests

- One sample test
  - One population, compare test statistic, e.g, sample mean, with a known number
- Two sample test
  - Two populations, compare two population means
  - Paired test
    - Two dependent groups, for example, same group been measured at two different times.
      - Essentially one sample test
  - Unpaired test
    - Two independent groups, may have different sample sizes.

# Two sample t-test

- Compares the means of the two groups of data
  - $X_1$ random sample from $N(\mu_1, \sigma_1^2)$

    $X_2$ random sample from $N(\mu_2, \sigma_2^2)$
  - $H_0 : \mu_2 = \mu_1 \quad H_a : \mu_2 != \mu_1$ (other $H_a$?)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{Var(\bar{x}_1 - \bar{x}_2)}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{Var(\bar{x}_1) + Var(\bar{x}_2)}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

# Student t test v.s. Welch t test

- If population variance from two samples are equal, use pooled variance (student t test)

$$t^* = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}, \; df = n_1 + n_2 - 2$$

- If population variance from two samples are not equal, use unpooled variance (Welch t test)

$$t^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{(n_1-1)\cdot(n_2-1)}{(n_2-1)C^2+(1-C)^2(n_1-1)} \qquad C = \frac{s_1^2/n_1}{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

A simplified way $df = \min(\mathrm{n1} - 1, \mathrm{n2} - 1)$

# Hypothesis Test - Interview Quiz

- Test a coin is fair. What test? What is null hypothesis?
- Test two group of users have same CTR (click through rate). What test? What is d.f. (degree of freedom)?
- Test two group of users on your website have the same mean spending. What test? What is d.f.?

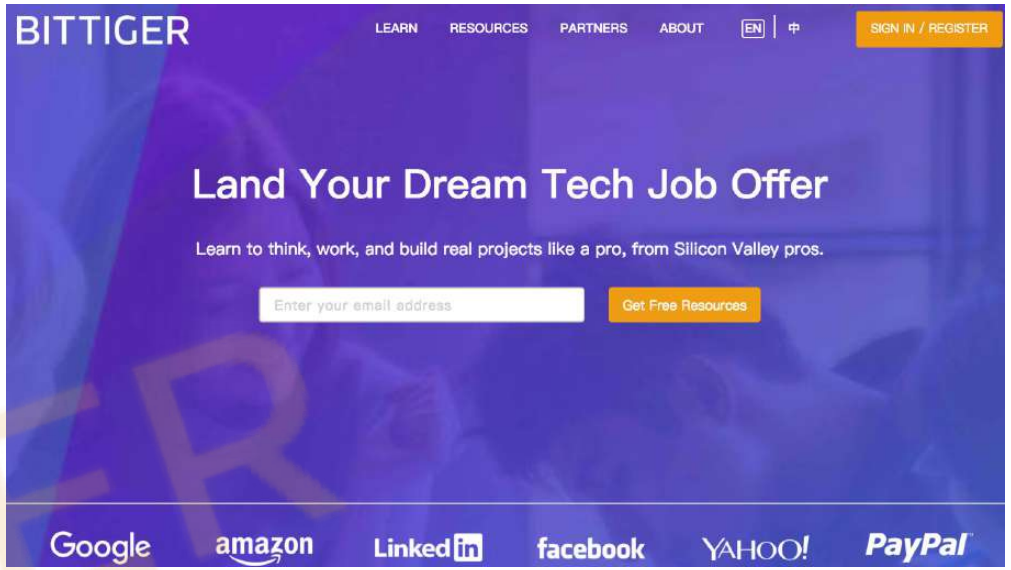
# Assumptions of t test

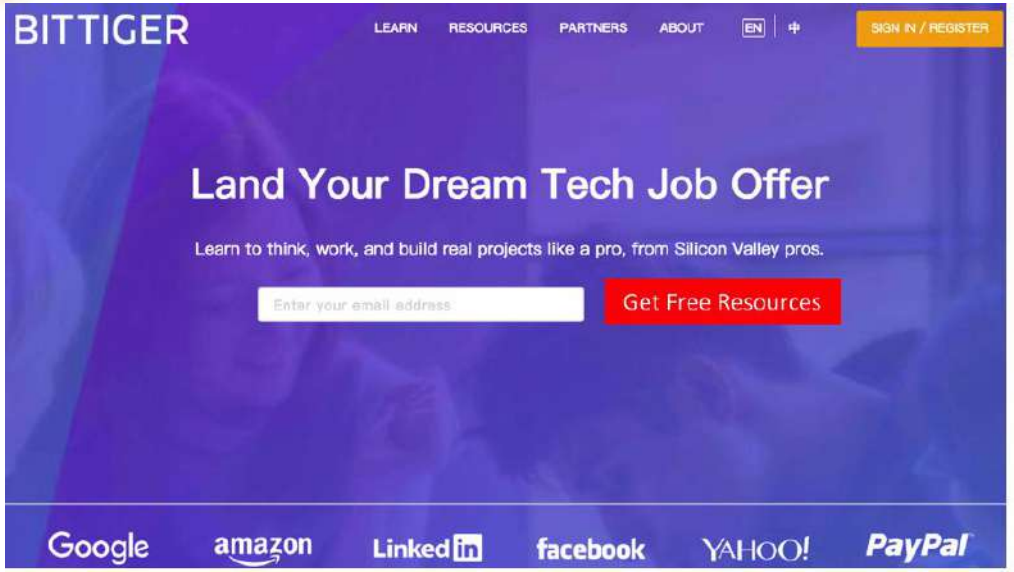
- Student t test
  - Normality
  - Independence
  - Equal variance (two sample test)
- What if normality is violated
  - Just do it (CLT)
  - Transformation
  - Other nonparametric methods

# Process of A/B Testing

# Process of A/B Test

- Design
  - Understand Problem & Objectives
  - Come Up with Hypothesis
  - Design of Experiment
- Implement
  - Code change & Testing
  - Run Experiment & Monitor
- Measurement
  - Result Measurement
  - Data Analysis
  - Decision Making



# Process of A/B Test

- Design
  - Understand Problem & Objectives
  - Come Up with Hypothesis
  - Design of Experiment
- Implement
  - Code change & Testing
  - Run Experiment & Monitor
- Measurement
  - Result Measurement
  - Data Analysis
  - Decision Making

# Design of Experiment

# Design of Experiment (DOE)

## Outline

- Key Assumptions
- Assignment
- Metrics
- Exposure & Duration
- Sample Size Calculation

# DOE - Key Assumptions

- The factor to test is the only reason for difference
- All other factors are comparable
- A unit been assigned to A or B is random
- Each experiment unit are independent

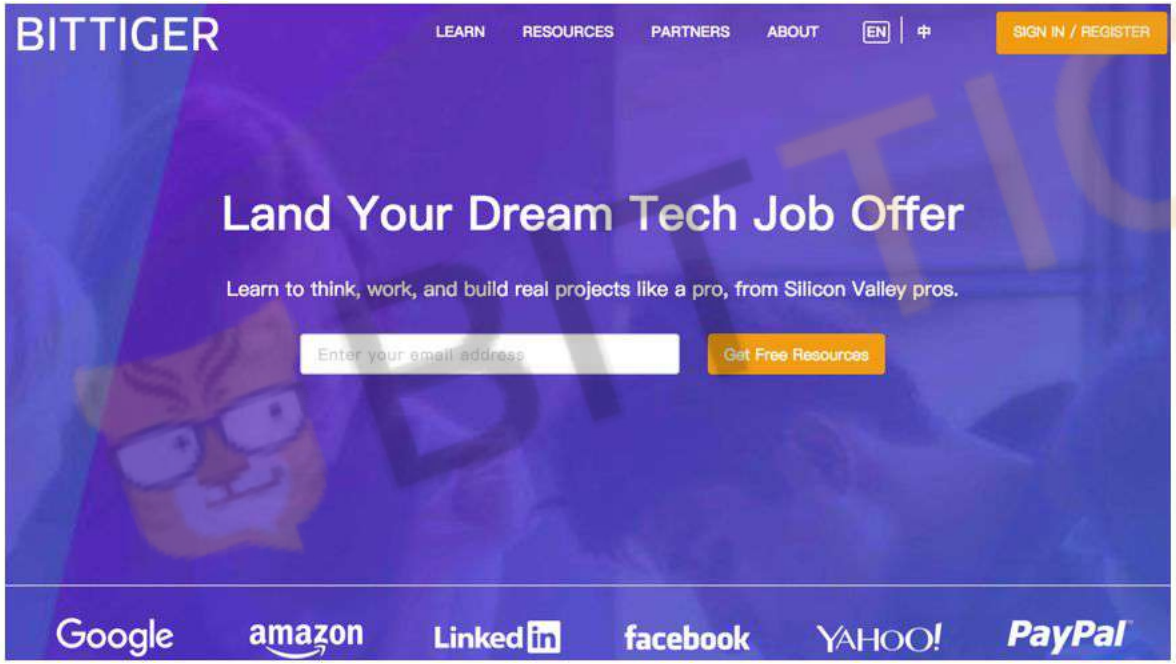
Principles of Experiment Design

- **Independent** samples
- **Block** what you can control
- **Randomize** what you can not control

# DOE - Assignment Unit

BitTiger
https://www.bittiger.io/

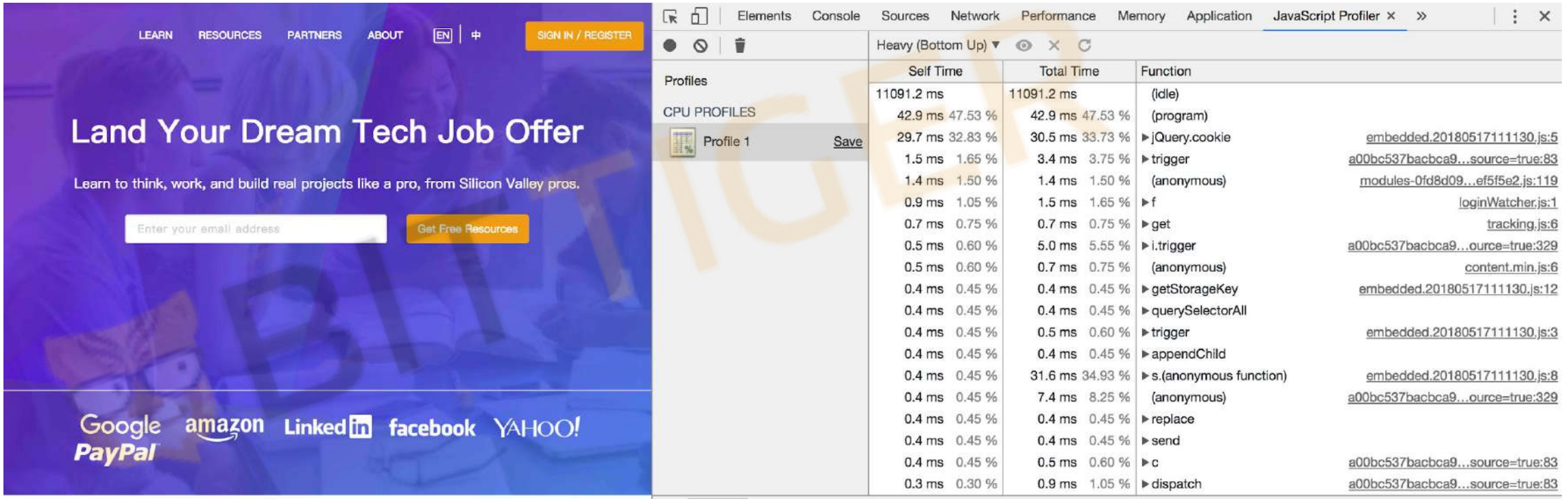
How to decide which version to display to whom?




# DOE - Assignment Unit

## What is the unit to split A/B?

User_id? Cookie_id? Device_id? Session_id?IP address?  etc



# DOE - Assignment Unit

## Considerations

- What are the eligible subjects we try to influence
  - Example 1: Pop up promotion ‘15% off if register today’
  - Example 2: Send emails with ‘new dress arrivals’
- What is the objective
  - Example 1: Remove homepage animation to reduce loading time
  - Example 2: Change button color to have more users to click
  - Example 3: Test impact of a change in ETA calculation method on trip cancellation rate. Rider_id, driver_id, trip_id?
  - Example 4: Test impact of a change in ETA calculation method on user retention rate. Rider_id, driver_id, trip_id?
- Independence & User experience
  - Example 1: Change homepage design in an app
  - Example 2: Add new video chat filters

# DOE - Assignment Unit

In Practice, assignment unit

- Default is user_id
- Sometimes there is not only one right answer. Have to make a decision but aware of the pros & cons

Split % - % of Users in test / control

- Most common 50/50 split
- Sometimes not
  - Time sensitive e.g. holiday marketing campaign

# DOE – A/A Test

- Randomly assigned
- Test / Control % is as designed
- One unit only in one group
- All other factors are comparable

**How to Check?**

A/A Test: use A/B test framework to test two identical versions against each other. There should be no difference between the two groups.

The goal:

- Make sure the framework been used is correct
- Data exploration & parameter estimation (e.g. sample variance)

# DOE – Assignment Common Problems

| Assumptions | Practical Problems | Potential Solution |
|---|---|---|
| Independence | Non login User | Assign by device_id, cookie_id, etc,<br>Predict user with models |
| | Multi device user | |
| | Multiple user share a device | |
| Assignment to T/C is random | Bug resulting in deterministic assignment | Assignment check by group, fix bug if identified |
| Reproducible | | Set Salt |
| 50 / 50 split (or other % as set) | Imbalance assignment due to experiment setup | Understand why, change assignment method |
| Test / Control are comparable in all dimensions except treatment factor | Pre-bias (run A/A test to check) | set a different salt,<br>Post experiment adjustment |

# DOE - A/A Test - Interview Quiz

Your colleague gave you two dataset and told you they are test/control groups from an A/B test. What will you do to make sure the datasets are appropriate to use?

What will you do if you find users been assigned to both test and control group? Do you have any concern?

# DOE - Metrics

What to compare?

How would you know the impact of your experiment

How would you make a business/product decision with your experiment?

We need metrics!

# DOE - Metrics

Metrics should be set before experiment start

- Understand what kind of changes your experiment would cause
  - Usually multiple changes happen the same time
  - Trade-offs
- Understand what are the metrics worth monitoring
- Understand the importance of these metrics
- Set expectation how these metrics would change

# DOE - Metrics

What are the potential positive & negative impact of following experiments?

- Example 1: Remove homepage animation to reduce loading time
- Example 2: Change button color to have more users to click
- Example 3: Use real time traffic data to make ETA calculation more accurate to increase matching efficiency
- Example 4: Add new video chat filters to increase user engagement

# DOE - Metrics

Set key evaluation metrics

- That is what you use to make a decision
- Usually one or a few
- Sometimes use a comprehensive evaluation metrics (e.g. weighted average of three metrics)

Other metrics worth monitoring

- Avoid unwanted negative impact
- Understand change in other metrics

# DOE - Metrics - Interview Quiz

What metrics are you going to evaluate the success of digital ads? How to convince clients to buy ads on our website?



# DOE – Exposure & Duration

## Should you show the A/B version to all users?

No. may cause bad user experience if test version is bad

Start with a small proportion, like 5%, gradually roll out to more users

## How long are you going to run your experiment?

In practice, we want to minimize the exposure and duration of an A/B test, because

- Optimize business performance as much as possible
- Potential negative user experience
- Inconsistent user experience
- Expensive to maintain multiple versions

# DOE – Exposure & Duration

How to decide exposure %?

- Size of eligible population
- Potential impact
  - User experience
  - Business impact
  - Easy to test & debug?

  - Example 1: Redesign the layout of your app. May significantly change user behavior. Needs three teams of engineers to coordinate
  - Example 2: Change button color to have more users to click. Need one engineer 10 mins to make a change

How to decide duration?

- Minimum sample size
- Daily volume & exposure %
- Seasonality (at least one seasonal period)

# Power Analysis & Sample Size Calculation

# Type I, II error, power

| | Ground Truth | |
|---|---|---|
| Decision | $H_0$ | $H_a$ |
| Not reject $H_0$ | Correctly (not reject null) | Type II error, β |
| Reject $H_0$ | Type I error, α | Correctly (reject null), power |

$$\alpha = P(reject\ H_0 \mid H_0)$$

$$\beta = P(not\ reject\ H_0 \mid H_a)$$

$$power = P(reject\ H_0 \mid H_a)$$

# Data Assumptions

- What distribution assumptions are you making to your data?

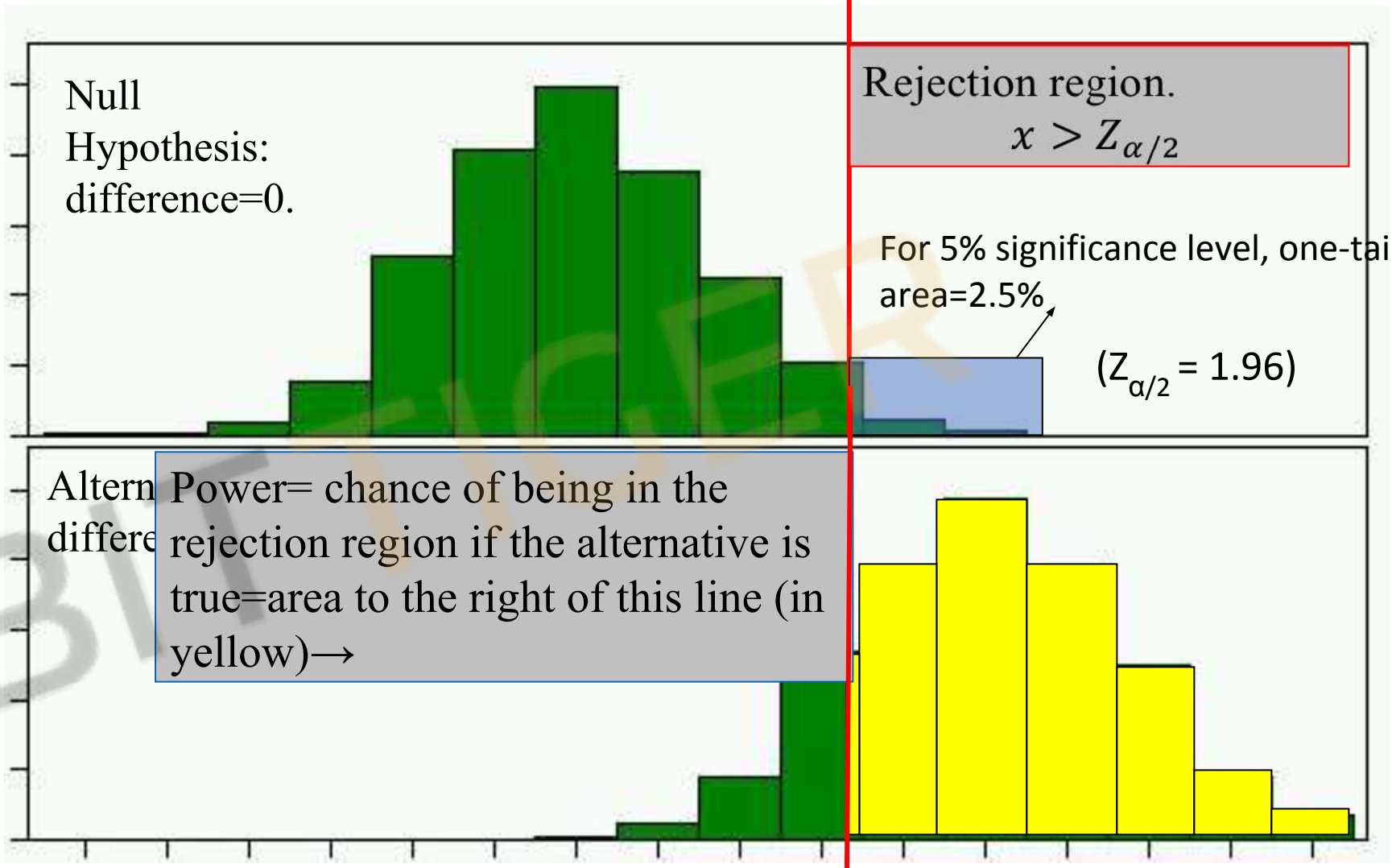
  i.i.d. Normal distribution, Central Limit Theorem

- What is the null hypothesis of your test?

$$diff = \mu_A - \mu_B = 0$$

# Power Analysis

$diff \sim N(0,1)$




$diff \sim N(3,1)$

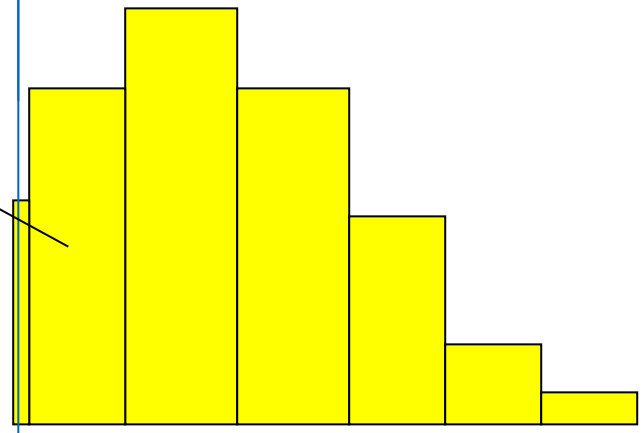
# Power Analysis

Rejection region.
Any value >= $Z_{\alpha/2}$

Power= chance of being in the rejection region if the alternative is true=area to the right of this line (in yellow)

Power here:

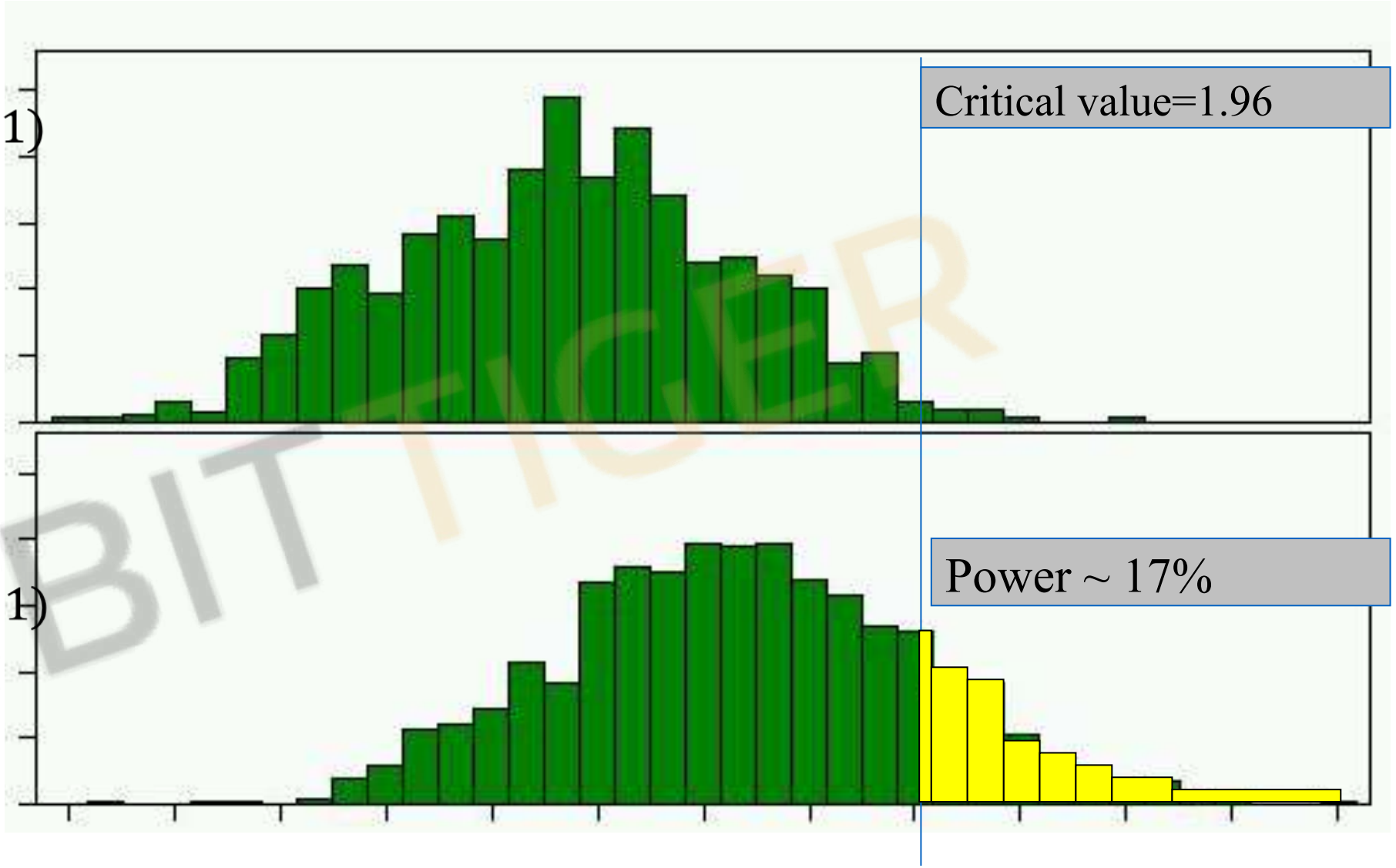
$$P(X > 1.96 \mid \mu = 3, \sigma = 1)$$
$$= P(Z > \frac{1.96 - 3}{1})$$
$$= 85\%$$



# Power Analysis

$diff \sim N(0,1)$

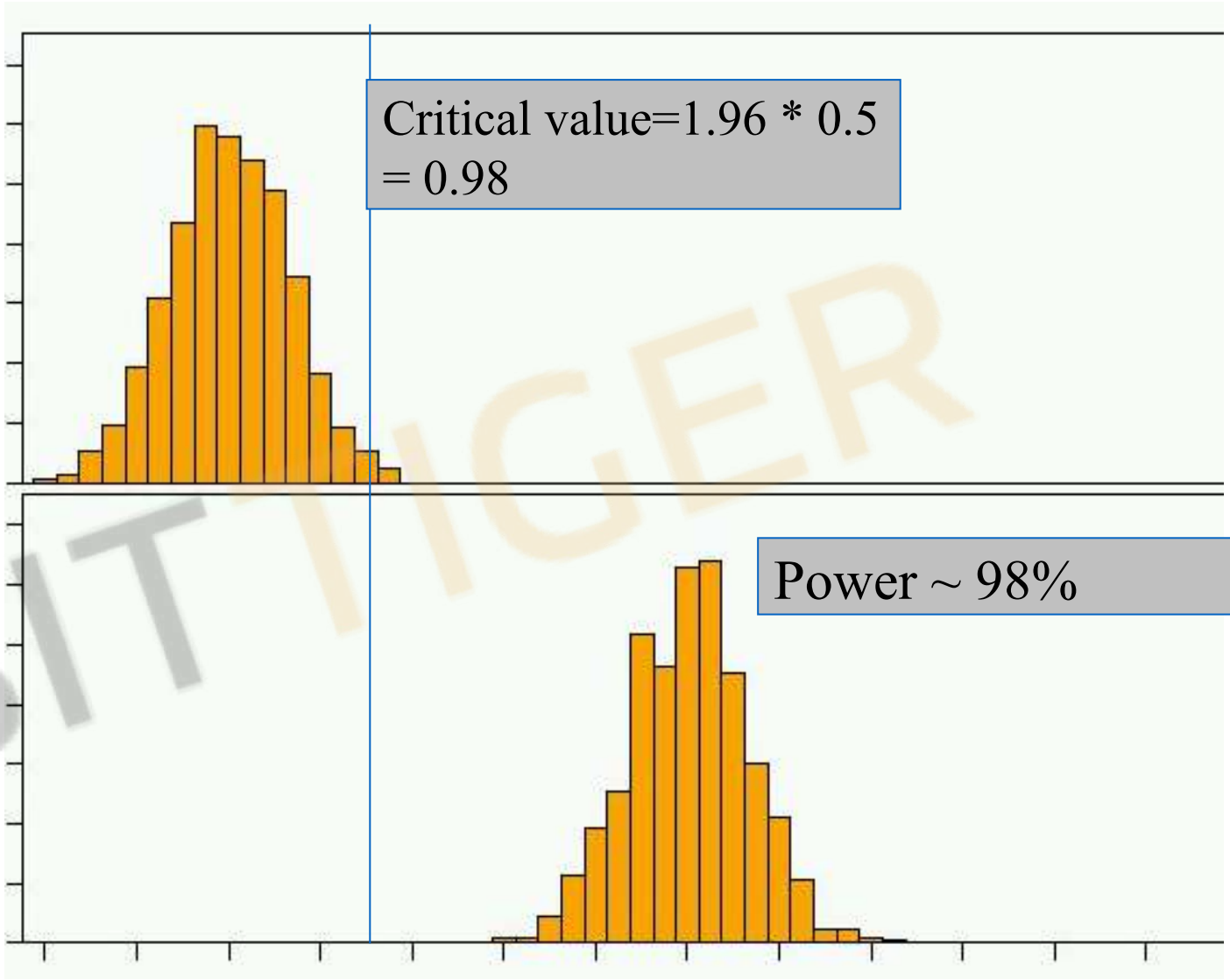
$diff \sim N(1,1)$



# Power Analysis

$diff \sim N(0, 0.5)$

$diff \sim N(2, 0.5)$



# Factors Impact Power

How is the power change if the following factors increase?

1. Size of the effect ↑
2. Variance of distribution ↓
3. Significance level desired $\alpha$ ↓

# Sample Size Calculation

Sample size in each group (assumes equal sized groups)

Represents the desired power (typically .84 for 80% power).

$$n = \frac{2\sigma^2 (Z_\beta + Z_{\alpha/2})^2}{\text{difference}^2}$$

Standard deviation of the outcome variable

Effect Size (the difference in means)

Represents the desired level of statistical significance (typically 1.96 for 95%).

# Sample Size Calculation

$For\ given\ \beta\ (power), \alpha(significance\ level), \sigma\ (standard\ diviation\ of\ data)$

$$Z_\beta = \frac{critical\ value - diff}{standard\ error(diff)} = \frac{z_{1-\alpha/2} * SE(diff) - diff}{SE(diff)}$$

$$= -Z_{\alpha/2} - \frac{\text{diff}}{\text{SE(diff)}} = -Z_{\alpha/2} - \frac{\text{diff}}{\sqrt{2\sigma^2/n}}$$

$$\therefore \text{n} = \frac{2\sigma^2\left(Z_\beta + Z_{\alpha/2}\right)^2}{diff^2}$$

$$\text{SE(diff)} = \sqrt{Var(diff)} = \sqrt{\text{Var}(\overline{X_a} - \overline{X_b})}$$

$$= \sqrt{\text{Var}(\overline{X_a}) + Var(\overline{X_b})}$$ as $\text{X}_\text{a}$ and $\text{X}_\text{b}$ are ind

$= \sqrt{2\sigma^2/n}$ n is sample size of one group, assuming two groups have equal sample size

If not equal variance, $\text{SE(diff)} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$

# DOE – Sample Size

$$n = \frac{2\sigma^2 (Z_\beta + Z_{\alpha/2})^2}{\text{difference}^2}$$

Need to Estimate:

- Variance - estimate with sample variance
- Difference – opportunity sizing
  - Observational data
  - Qualitative result (e.g. survey, small scale test)
  - Intuition & Practical consideration (e.g. minimum effect worthing implement the change)

# DOE – Sample Size - Interview Quiz

How long would you run your experiment?

What is your minimum sample size? What factors would you consider? How would these factors impact your sample size?

What will you do to balance user experience and quick learning?

What is your roll-out plan?

What will you do if your experiment take too long to run?

# Implementation

# DOE – Peeking

Why calculate sample size?
Can we just let the experiment run until the result is statistically significant?

No! Highly increase false positive rate

Type I error (false positive) $\alpha$ = 0.05 ->

When null hypothesis is true, the chance of reject H0 is 0.05

What is the chance of seeing at least one rejection having 10 tests simutaneously?

$$1 - (1 - 0.05)^{10} = 0.4$$

# DOE – Monitoring

Monitor key metrics while experiment running

- Should **NOT** frequently check result
- Should **NOT** stop once result turns significant
- Wait until get minimum sample size from experiment design
- But need to monitor for alarming changes. Pause and investigate if needed

# DOE – Problems & Solutions

1. What if it takes too long to get desired sample size?
    - Increase exposure
    - Reduce variance to reduce required sample size
        - Blocking – run experiment within sub-groups
        - **Propensity Score Matching**

Example:

We want to test the impact of a product change on ads' click through rate.

We know there are users who are more clicky and have higher CTR in general and users who almost never click on an ad. Is it fair to compare all users directly?

# Propensity Score Matching

## Procedure

1. Run a model to predict Y (CTR rate) with appropriate covariates

   Obtain propensity score: predicted y_hat

2. Check that propensity score is balanced across test and control groups
3. Match each test unit to one or more controls on propensity score:
   - Nearest neighbor matching
   - Matching with certain width
4. Run experiment on matched samples
5. Conduct post experiment analysis on matched samples

# DOE – Problems & Solutions

## 2. What if your data is highly skewed or statistics is hard to approximate with CLT?

Example:

1. Metrics like revenue is highly skewed & have outliers
2. In risk/fraud, most transactions have no loss while some fraud transactions have very high loss

### Solutions

- Transformation (hard to interpret)
- Winsorization / Capping
- **Bootstrap**

# Bootstrap

Bootstrap is a resampling method. It can be used to estimate sampling distribution of any statistics, commonly used in estimating CI & p-value & statistics with complex or no close-form estimator

## Procedure

1. Randomly generate a sample of size n with replacement from the original data. n is the # of observations in original data
2. Repeat step 1 many times
3. Estimate statistics with sampling statistics of the generated samples

Practice:

Use R to generate a 100 sample from Normal(3, 5).

Calculate it's theoretical & bootstrap estimate of mean & variance

# Bootstrap

## Pros

- No assumptions on distribution of original data
- Simple to implement
- Can be used for all kinds of statistics

## Cons

Computational expensive

# Interview Quiz

- Can you run an experiment and keep reading until the result is significant?
- What if your key metrics dropped by 5% on first day? What if dropped by 20%?
- What is bootstrap? Boosting?
- The metrics of interest is 90th quantile of users' spending. How to estimate sample mean and variance?

# Result Measurement

# Result Measurement

- Data Exploration
  - Imbalance Assignment
  - Mixed Assignment
  - Sanity Check
- Hypothesis Test
  - Conduct test
  - Multiple Testing
- Result Analysis
  - Pre-bias Adjustment
  - Analysis unit different with Assignment Unit
  - Cohort Analysis

# More Advanced Topics

# RM – Data Exploration

- Data Exploration
  - Check for % of test/control units. Is the % matching DOE?
    - IF not match, need to figure out what's the cause
  - Check for mixed assignment
    - It's hard to resolve. If # of mixed samples is small, OK to remove. If big, need to figure out what's the cause
    - What's the problem of throwing away mixed samples?
  - Sanity Check
    - Are test/control similar in other factors other than treatment?

# RM – Hypothesis Test

- Set up the right test
  - Mostly use T-test
  - When variance is known is large, can use Z-test
  - When sample size small can use non-parametric methods
  - For complicated statistics, can use bootstrap to calculate p-value

# RM – Decision Making

- If all metrics move positively
  - Meet expectations? Yes, ready to launch
  - Be cautious if result is too good. May need to investigate (e.g. outliers)
- If some metrics move negatively
  - Are they as expected? Are these metrics important?
  - Deep dive to find causes
- If results are neutral
  - Slice / Dice on sub-groups

# RM – Interview Question

- What if your result show positive impact on some metrics and negative impact on some other metrics?
- What if your result is neutral?
- What if you result is statistically significant but the margin is very small?
- Take home challenge

# Multiple Testing

What if you have multiple test groups?


| | Image | Headline |
|---|---|---|
| VERSION 1 | | "ACME WIDGETS" |
| VERSION 2 | | "ACME WIDGETS" |
| VERSION 3 | | "THE ONE AND ONLY ACME WIDGETS" |
| VERSION 4 | | "THE ONE AND ONLY ACME WIDGETS" |

False positive rate is much higher when doing multiple testings!!
Need to control family-wise false positive rate

https://www.optimizely.com/optimization-glossary/multivariate-test-vs-ab-test/#

# Multiple Testing Adjustment

**Bonferroni Adjustment**:

Assume we have m tests, Set $\alpha_i = \frac{\alpha}{m}$ for each experiment.

This guarantee the overall FP < $\alpha$, but too conservative

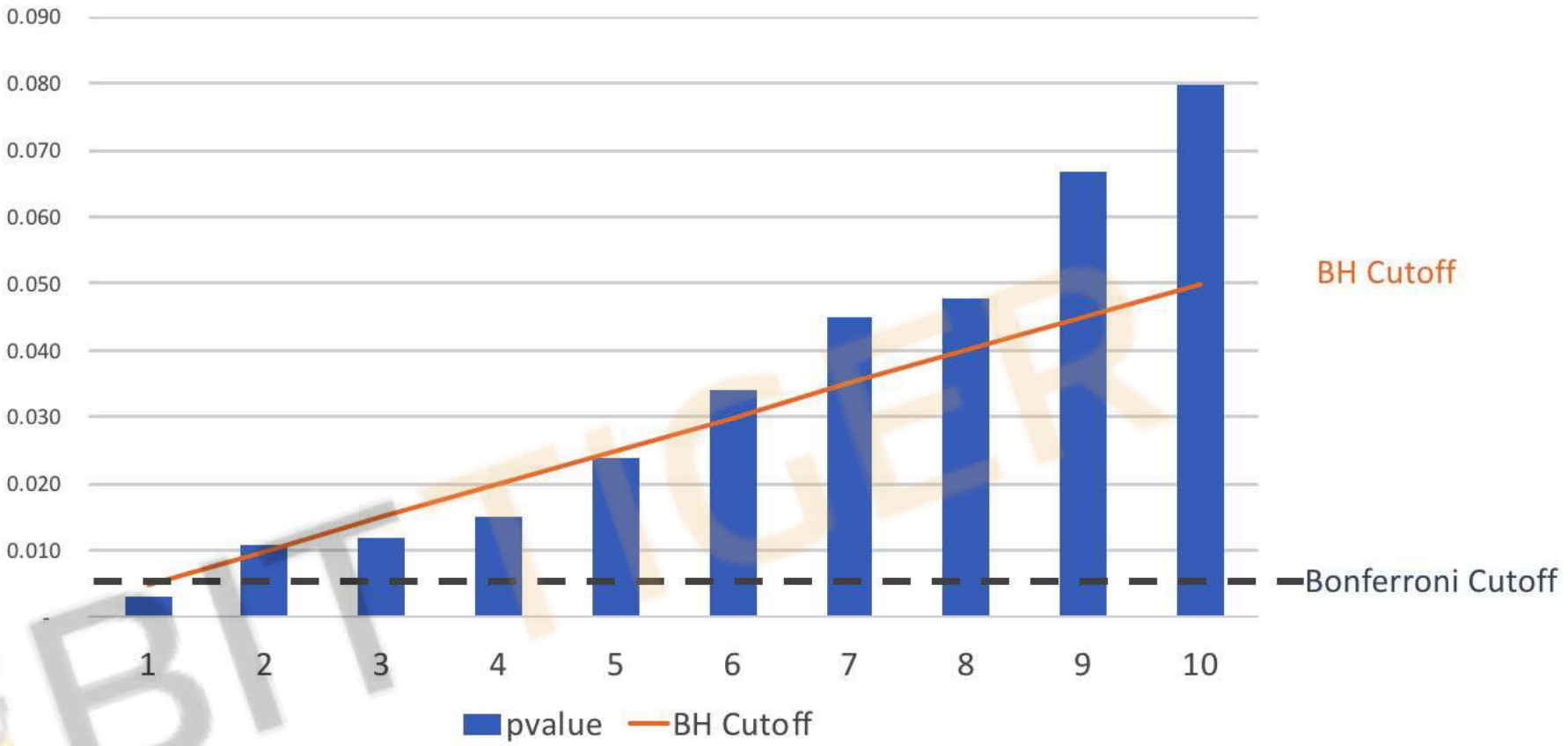
# Multiple Testing Adjustment

**FDR (false discovery rate) Adjustments**

Benjamini – Hochberg Adjustment:

1. Rank p-values $P_i$ of m tests from low to high
2. Find the largest k such that $p_k \leq \frac{k}{m} * \alpha$
3. Reject experiments 1,…,k. Accept experiments k+1,…, m

# Multiple Testing Adjustment




Bonferroni Method: Reject Test 1, accept all rests

BH Method: Find max k that $p_k \leq \frac{k}{m} * \alpha$ , for this case, k = 5. Reject T1 to T5, accept T6 to T10

# Multiple Testing Interview

- 第一个人 组里比较 senior的感觉 给我介绍了一个他们之前做的问题大概就是Neyman Rubin 模型 但是记录的output有很多 要判断这些output中有没有任何一个的均值是明显有差异的 也就是一个multiple testing的问题 用 Bonferroni correction。然后问我实际发现correct之后没有一个p-value比threshold小 但是很多很接近 那么有没有办法来处理。
- 第二个人 介绍了另外一个case 就是在做ab testing的时候有可能会treatment组 出错 所以希望尽量避免这种情况 但是另一方面 又希望可以把尽量多的人放进去 如何formulate这个trade off （type1 type2 error）

# Pre-bias Adjustment

We had the assumption that the A/B groups have no difference before experiment. What if the does exist difference?

**Regression Adjustment**

Y_post = beta_pre * Y_pre +  beta_t* Treatment_Group

**Diff - in – Diff Comparison**

(Y_post/t – Y_pre/t) – (Y_post/c – Y_pre/c)

# Cohort Analysis

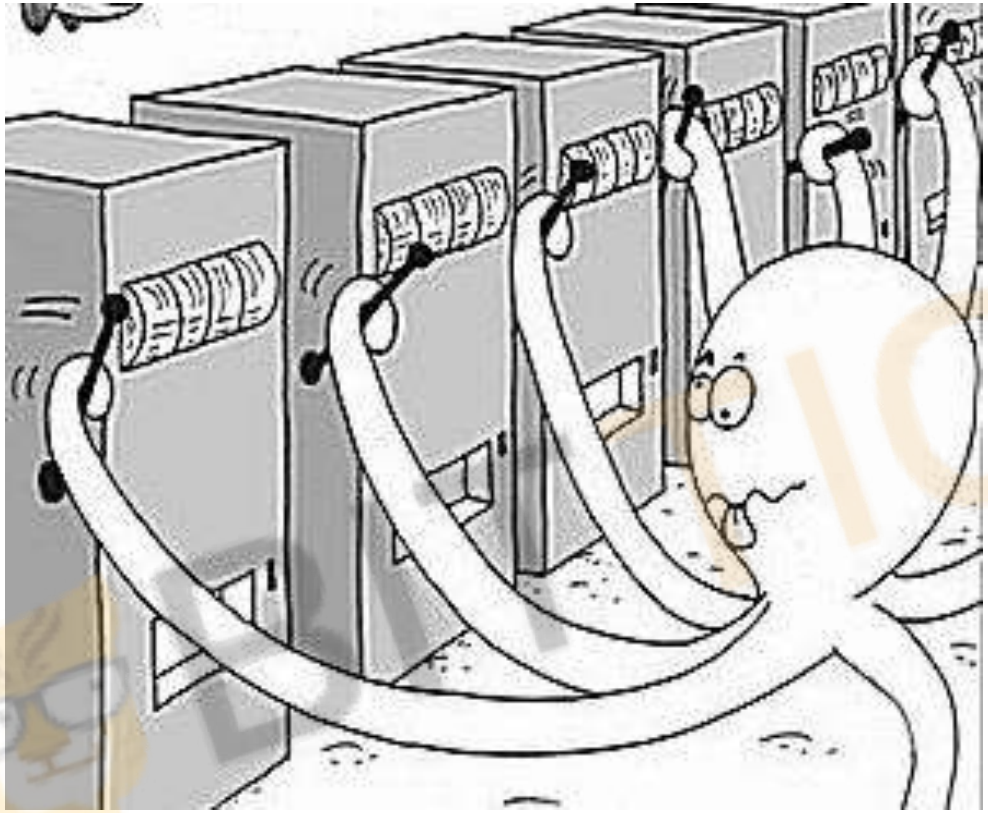
## How to measure impact over time?

E.x. Spotify is testing a new recommendation system algorithm, which is expected to give more accurate recommendations thus improve user engagement. You do not expect users to notice the difference and take any actions since day 1. But users are expected to gradually get more engaged over time

**Cohort Analysis**

Select a cohort of users (e.g. T/C users assigned on the same day) and monitor their metrics change over time

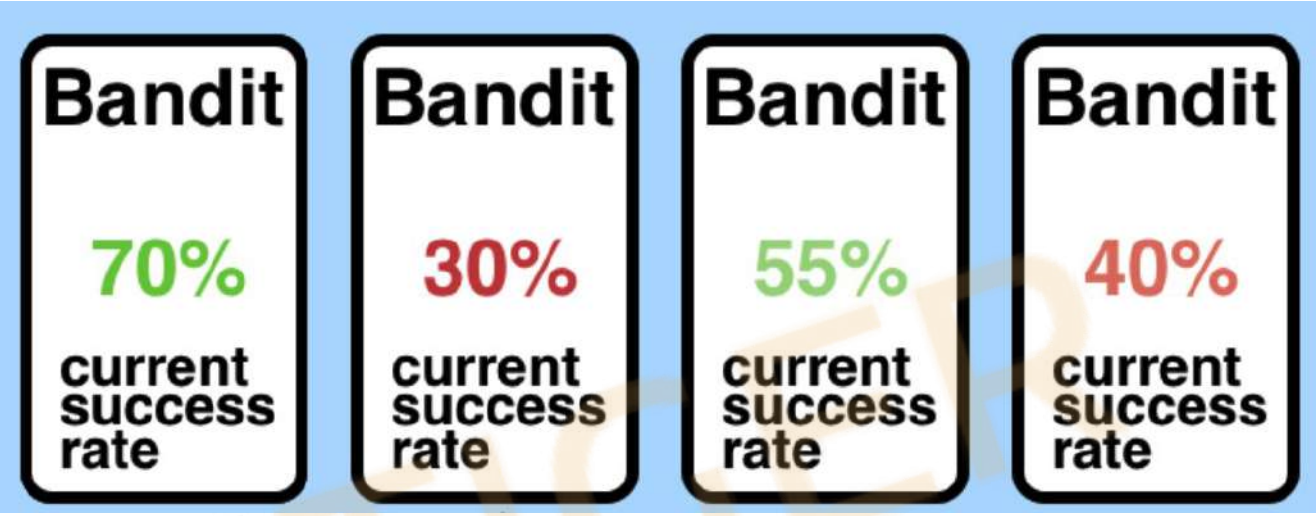
# Multi-arm Bandit Problem



- Each slot machine has different rewards
- Objective: to maximize rewards in casino

Which arm would you pull?

# Multi-arm Bandit Problem

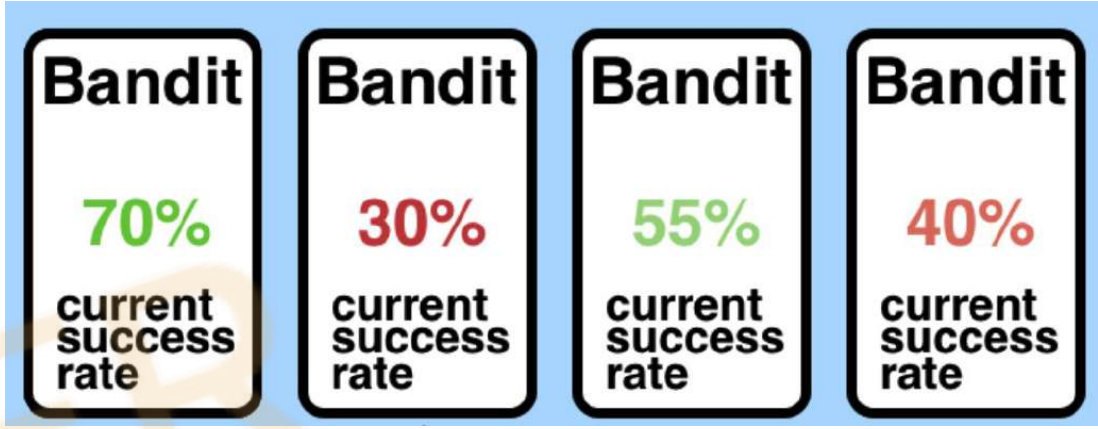



- You have multiple test groups. Each group has different success rate and are unknown before your experiment. Your goal is to maximize the overall success rate of all users. How would you allocate your users?

# Multi-arm Bandit Problem

Two stage

- **Exploration**: Example Split 10% of all of your users equally between all of the treatments
- **Exploitation**: Example Use the other 90% of the tokens in the machine that rewards you the most




**Different Methods:**

**Epsilon-Greedy**: the rates of exploration and exploitation are fixed

**Upper Confidence Bound**: the rates of exploration and exploitation are dynamically updated with respect to the rewards of each arm

**Thompson sampling**: the rates of exploration and exploitation are dynamically updated with respect to the entire probability distribution of each arm

https://www.searchenginepeople.com/blog/16072-multi-armed-bandits-ab-testing-makes-money.html

# Limitations of A/B Test

- Highly rely on your hypothesis
- Good for optimize small changes, Not good for innovative changes, long term strategies
- Other factors involved: e.g. learning effect, network effect

Other ways to make product improvement

- Qualitative Studies – Survey, focus group
- Observational studies

# Typical Interview Examples

# A/B Test in Interviews

## How are A/B test questions asked?

- Asked directly
- Asked in case studies / product sense questions (Most common)
- Asked in take home projects

# Example Question 1

Survey showed teenagers are less engaged with Facebook after their parents join FB. What to do?

1. Understand problem, define population & objectives & metrics
2. Brainstorm features to consider
3. How do you know if your feature works?

- Design of experiment (metrics, assignment)
- Duration & exposure of your experiment
- How would you make decision?
- What if you see xyz?

# Example Question 3

You have 1M budget to spend on holiday campaigns. Possible investments include mailed ads / emails / display ads / search engine / social media ads. How would you optimize the budget?

1. Define objectives & metrics
2. Design of experiment
3. Result measurement
4. Slice / dice analysis
5. Multi-armed bandit

# Example Question 3

An engineer suggest promoting new sellers on your website to boost seller growth. But another engineer is worried about it hurting overall sales. How would you make a decision?

1. Define objectives & metrics
2. Target metrics vs monitor metrics
3. Result measurement
4. Decision making with mixed results

# Example Question from students - 1

1.slow roll out: How do you explain the result

**Table 1: Conversion Rate for two days.**
**Each day has 1M customers, and the Treatment (T) is better than Control (C) on each day, yet worse overall**

| | Friday<br>C/T split: 99%/1% | Saturday<br>C/T split: 50%/50% | Total |
|---|---|---|---|
| C | $\frac{20{,}000}{990{,}000} = 2.02\%$ | $\frac{5{,}000}{500{,}000} = 1.00\%$ | $\frac{25{,}000}{1{,}490{,}000} = 1.68\%$ |
| T | $\frac{230}{10{,}000} = 2.30\%$ | $\frac{6{,}000}{500{,}000} = 1.20\%$ | $\frac{6{,}230}{510{,}000} = 1.20\%$ |

# Example Question from students - 2

2.Interpret AB testing result, Treatment effect for each group as below:

- Trt1: -5, CI: (-7.5, -2.5)
- Trt2: -15, CI: (-17, -13)
- Trt3: -12, CI: (-28, -4)

How to interpret C.I.? Which treatment to choose? Increase test power / accuracy?

# 1-1 Conversation in Comments

In the earliest version of Facebook, it only has function of post and comment, comments are affiliated to post. The dataset contains only user_id and timestamp of user actions. how to identify whether conversations happen under post. conversation definition / 如果有一个简化版的FB，只有简单的文字post和comment且comment 没有nested这个功能（只能一条一条）。问如何判断一个post的comment包含conversation。有什么metrics可以监测。

1) Ask for assumption and give my own assumption: assume that I have the data with user_id, post_id, and timestamp. Is the table on the comment_id level? That is, there are unique user_id and unique timestamp for each comment.
2) Define what is a conversation: at least one person inside the conversation should have at least 2 comments. The pattern should look like ABA, ABAB, ABACA.
3) Algorithm (key points - back and forth pattern and time interval):

- Keep the users who have at least 2 comments do that we can find the first person who potentially in a conversation.
- For user one, assume the time interval of his/her first and last comment is between A and B, keep the users who have at least some comments between A and B. For user two that may have a conversation with A, there must be one comment between time A and B.
- To find a third person in this conversation, he/she should have at least one comment between time interval of user one or the time interval of user two. By repeating this, we can find a group of users who may have conversation under the post. If there is no such a group, then there may not be any existing conversation
- The basic idea is that there are users leaving multiple comments and there is a back-and-forth user_id pattern if ordered by timestamp

4) 可能有什么样的false positives
5) 如何计算likelihood of a post to have at least 1 conversation?
   logistic regression with training data (we can read some comments by person and define whether there is any conversation). Use metrics to estimate the probability:
   - The ratio of #comments over #user_id for each post
   - The ratio of # comment over frequency of leaving comments for each user (同一用户的评论数/评论频率)
   - 连续评论的人的mutual friends
   - 用户交替评论的pattern (有没有ABAB这样的连续两个人评论);
   - 评论时间间隔
   - Category of the posts
6) To use the metrics, I may try to read some posts and determine if there are any conversations. For example, I can read 100 posts and calculate the ratio between comments/user, and can take the lowest value from the ones do have a conversation as the threshold. Then I can determine whether my metric is big enough.

# 1-2 Comments

1) how to measure health of facebook
   - website/sytem本身health，是不是function normally
   - user engagement (新用户注册率，现有用户的interactions包括comment/share blablabla，然后DAU/MAU之类的)
   - retation rate

- revenue (AARRR)

2) focus在comment上，如果要求你做一个dashboard specifically about comment，有哪些metrics你可以present
   - show distribution of comments/users
   - average comments per post/user，numerator等于总评论数，denominator是distinct users who have left at least 1 comment per day
3) 现在把denominator换一换，改成DAU，我们发现这个总comment/DAU的比率跟一年前同一天相比，增加了50%,有哪些可能原因
   - data accuracy? Spams or outliers?
   - comments increase or DAU decrease? probabily comments
   - any UI design change or other experiment on comments? 有新feature刺激用户留言更多之类的
   - slice users into differenct categories: user mix可能有变化，可能今年新增的growth都是愿意发更多comment的用户
   - content内容的变化吸引更多user leave comment
4) 如何判断是哪种?
   - segmentation by users' language/platform/device type/browser/location看看有没有哪个subgroups特别突出，有的话可以dive deep。

# 2. fb要在messenger里加一个类似微信支付的功能，你可以转账给别人也可以付费

1) 这样的好处坏处;
   - To FB

A: it can bring some new users because if many of my friends are using it to transfer money, I will also register and install the APP to user it
A: active more users because many users would need this feature and they will user it to transfer money to their friends. If more users are aware of this feature, they will be active in messenger APP
R: users would feel messenger is more useful and they will occasionally use this feature
R: if I like this feature and user it a lot, I will referral it to my friend and invite them to use it, so that the level of referral would also increase
R: there can be some service fee; or interest revenue because it takes time for the money transfer; there can also be some online financial products that the user can purchase
Cons: need to work with bank or third-party financial services; need to work more on how to protect user's privacy and safety

   - To users
     - Can transfer money more easily
     - Worries of privacy and safety

2) 问说有什么framework可以evaluate这个feature吗?
   - # of new users for messenger & for the new feature
   - number of transactions/ratio of users who use the feature
   - Conversion rate: click the transfer icon – add bank/card info – find the friend to transfer – enter the amount – finish transfer
   - Retention rate
   - How many new users acquired by referral
   - Revenue (ROI)

3) estimate distribution of number of payments in a month (x轴是number of payments, y是number of users); 求这个distribution的mean和median, 是不是会有偏差?

- Exponential distribution: frequency of having zero or just several-time transaction would be pretty large.
- Poisson distribution: the probability of a given number of events occurring in a fixed interval of time or space (if these events occur with a known constant mean rate and independently of the time since the last event)
- Median: if more than 50% of users use zero times, then it would be zero; if it's smaller, then find the Poisson quantile of (0.5-p)
- Mean: Poisson distribution for number of payments larger than zero (assume we have the data on the time interval of two payments of one user, e.g. three days, that is, the lambda is 10, which is 1 over 1/10 months. The parameter of Poisson distribution is lambda times t, which is 10 here. The mean should be lambda + 1 for users with non-zero payments. The mean for all users is p*0+(1-p)*(lambda + 1)

# 3. FB's user confirmation

1) Why FB ask user to verify email/phone
   - Helping you log in. If you forget your password, you'll need an updated mobile number or email address to reset it.
   - Suggesting People You May Know so you can connect with them on Facebook.
   - Showing you relevant ads. However, we don't ever sell personal information, including your mobile phone number, to anyone.
   - Helping keep your account safe through opt-in features like text message or email alerts for unrecognized logins or two-factor authentication.
   - Prevent spam/bot users
   - Double check to make sure the email/phone is correct

2) Pros and Cons of using email and phone

Email:

- hardly change; less worried about privacy
- easily being spam; lower rate of checking; in some countries, people do not use it often

Phone:

- easily access (most people have one and bring the phone all the time); easy verification (no need to login to the email inbox)
- difficult to access when go abroad or change number; worry about spam or privacy;

3) If one day FB realizes that the verification rate for phone number suddenly decreased a lot, why? (here strictly for new user verification purpose)

- data accuracy
- technical error
- spam user (tried similar phone number a lot in one day)
- slice: specific time/season, country/language, versions, platforms

# 4. 如果instagram 出了一个新的feature可以让用户在一个device上随意切换账号，

问怎么识别几个账号来自同一用户，做了实验后发现numebr of account increased，但是avg time spent没有变化问为什么， 然后问了是否要launch这个feature。/ Instagram为了使用户方便launch了可以迅速切换账号的button，以前想要切换必须退出当前账号再登陆，现在简化了这个步骤，点屏幕右下方就可直接切换 / instagram now have feature let users to switch accounts with one button

1) how to identify multiple accounts belonging to the same user?
   - Same device id, ip address, registration phone number
   - Similar user name
   - Shared friends, all following some accounts
2) total time spent flats and number of accounts increases. What are you hypothesis about why this happened? what data you need and how to testify your hypothesis?
   - novelty effect，一开始推出这个feature大家觉得有趣就去多建了几个账号但是the way ppl interact with Instagram并没有变化所以avg time spent没有变。我觉得是因为人们没有足够的精力细致的管理多个账户，或者对于使用多个账户还在学习中。
   - data是否correct? test有没有错，有没有population selection bias – AA test
   - cohort analysis去filter out用这个feature的用户，然后dive further to see what's going on.
3) 不做test直接就roll out 给所有用户好不好？不好，太risky，可能有用户会抵制，从而导致用户流失。
4) 哪些用户会抵制？哪些用户会不喜欢这个feature？父母会抵制，会在生活中制止刷子女花很多时间刷ins/people who don't like changes
5) same assumption as 2, how do you determine if this feature is successful launch? No ab testing at all since the feature already launched.
   - 问清楚这个feature的goal然后做cost-benefit analysis。结合opportunity size以及你想的一些metrics来决定是否有practical impact，impact有多大来决定。

* 请问如何判断这些账户来自同一个user?
我的回答有：device number，ip address，geographic data，registration phone number, following和follower的intersection (shared friends, shared follower)，还可以通过取user_name 的pattern

* demographic data，很快就被面试官抓住了漏洞，他就问我怎么用demographic 区分，我想了一想发现不好区分，因为使用这个功能的用户往往想获取不同的内容，大概率想在另一个账户上be anonymous，我直接表达了我的意思，他接着问有没有人不想be anonmous的？（楼主这个时候已经是抓耳挠腮了，原本以为这个面试更偏向考logic，没想到问的这么细节）对于这个follow-up我给的回答是：有一些人会为了给某个post点赞去创建很多账号，这样他们的目的就并不是获取内容，而是做popularity contest了。接下来又有follow-up：你怎么判断哪些账户属于该类?
我：可以根据点赞的时间间隔，往往这种点赞的时间间隔会更加evenly distributed，因为用户想一次性点完所有赞就会连续的切换账号。说完这个我就陷入了沉思，好在面试官给了一个提示，说我们有user name，follower，following的data）

我：
follow-up：follower会有什么intersection?
我：因为是水军账户所以follower不会有什么real people，可能就是一些广告，推销或者色情的账户

# 5. Instagram Checkout

1) 为什么要推出这个feature? 有什么好，要从fb和商户两个方面来讲

- Facebook: AARRR
- Business: large size of user (1B MAU); especially good and convenient for Early-stage online retailers, E-commerce newbies, Retailers with limited products, Established Instagram-first retailers, and Retailers with a lot of social traffic because (i) they do not need to own thier e-commerce platform; (ii) they don't have the revenue to create a custom shopping cart or pay for a full-scale solution; (iii) link to your product every time you post

2) 有什么坏处，有谁会不喜欢这个feature
   - Customer: do not like Ads & sponsored contents; worried about privacy & safety; believe and prefer brand identity
   - Instagram: how to handle large transaction / limited-inventory drops / prevent bots and resellers
   - Business: diverse category of goods; lose customer data; already have their own inventory, payment and shipping system

3) 怎么样去改善这个feature
   - Goal & any feedback/complaint/reports
   - Find a way to differentiate customers who like the feature & do not; find a way to track inventory in a more systematic way

4) 怎么看我们是不是应该推出这个feature
   - AB test
   - Cost-benefit analysis / ROI analysis
   - Qualitative research on user's experience

6. **fb group**功能如何衡量表现，以及需要哪些数据来选择是focus在小group还是大group(小group就是人数比较少的那种)，你认为哪种更值得被投资和继续发展

1) How to measure health of large/small FB Group
   - key actions (post/reaction) per group users都可以用

A: new registered users, new groups
A: Active users using FB groups, active groups number, percentage of users that are using group, time spend
R: User retention(active time), churn rate (how many users are no longer active over all users), resurrection (how many users churned and came back)
R: how many users are accquaired by referral
R: revenue, ROI

- 大group还会monitor group membership comparing to last month

2) If FB Group is so good that a lot people start to only use it but no other FB products such as story, what would you think?
   - Whether there are any overlaps between the two features and how people user it. Analyze user's behavior and characteristics of the two product, and

determine what are some factors that may influence the probability of using the two

- Compare the cost and benefit/ROI of the two, especially in the long-term. Also compare the level of AARRR for the two
- Try to link the two products together

3） We found posts in Group normally get more comment than regular post, why? How do you verify your hypothesis?

- More relevant content: recruit some users in an experiment and ask them to post relevant and irrelevant posts in the same group, and they compare the number of the comments
- More close relationship/networks: recruit some users in an experiment and ask them to post the same posts in the group, as well as in the regular section. Compare the number of the comments

4） In groups, we have large and small groups. We want to add features to one of them due to the engineering resources. Which group will you choose?

Large groups, small group may have bias due to users in small groups are more active. Larger group can avoid this bias. A/B test, assumption

# 7-1 怎么识别fake news

- Three types of fake news:

1) Serious fabrications (i.e., news items about false and non-existing events or information such as celebrity gossip)
2) Hoaxes (i.e., providing false information via, for example, social media with the intention to be picked up by traditional news websites)
3) satire (i.e., humorous news items that mimic genuine news but contain irony and absurdity)

- Assumptions

1) The linguistic approach attempts to identify text properties, such as writing style and content, that can help to discriminate real from fake news articles. The underlying assumption for this approach is that linguistic behaviors such as punctuation usage, word type choices, part-of-speech tags, and emotional valence of a text are rather involuntary and therefore outside of the author's control, thus revealing important insights into the nature of the text. The linguistic approach has yielded promising results in differentiating satire from real news (Rubin et al., 2016).
2) Fact-checking approaches rely on automated verification of propositions made in the news articles (e.g., "Barack Obama assumed office on a Tuesday") to assess the truthfulness of their claims (use external sources, such as fact-check web/database or social media). But external source can be unavailable, especially for latest news.
3) Fake news is different from deceptive content, which has explored domains such as forums, consumer reviews websites, online advertising, online dating, and crowdfunding platforms Fake news producers usually seek political or financial gain as well as self-promotion while deceivers have motivations that are more socially driven such as self-protection, conflict or harm avoidance, impression management or identity concealment. Second, they differ significantly in their target and in the form they propagate: fake news items are usually disseminated at larger scale through the Internet and social media whereas deception is more specifically targeted at

individuals. However, since both tasks deal with deceptive content, we hypothesize that there are linguistic aspects that might be shared between these tasks.

- Approaches

1) Linguistic features + machine learning approaches (first find some real news and fake news by hand, and then use SVM classifier):
    - Punctuation
    - Psycholinguistic features (the proportions of words that fall into psycholinguistic categories. LIWC is based on large lexicons of word categories that represent psycholinguistic processes (e.g., positive emotions, perceptual processes), summary categories, as well as part-of-speech categories. Categorized into summary categories (e.g., analytical thinking, emotional tone), linguistic processes (e.g., function words, pronouns), and psychological processes (e.g., affective processes, social processes)
    - Readability. We also extract features that indicate text understandability. These include content features such as the number of characters, complex words, long words, number of syllables, word types, and number of paragraphs, among others content features.
    - Syntax
2) Check by id: Key with fraud is, not happening only once. People who commit fraud would like to repeat it if not being caught. all variables are really about something that should be unique but is not or extreme values. Hence two main ways to capture fraud:

i. Same device IP/Bank account/phone number as existing accounts;
ii. Anomaly detection-find outliers (extremely low price)

Ø More specifically with market place posting, we can address the listing and seller. For listing, pictures cannot be stolen from elsewhere/descriptions cannot be copied/resolution should not be too low/price should not be two

Ø With fake profile (say fake school): using ML algorithm or anomaly detection to find outliers. For instance, you may include the percentage of connections went to the same school/interaction with people from the same school/acceptance rate for the same school request as variables. In order to minimize the fake profile, you may want to use 2-step verification for risky users (minimize false negative you may not apply this to all users).

3) Use the existing fact-check database
    - Accuracy check

1) SVM: accuracy, precision, recall, and F-score as performance metrics
2) The rate of recognized spam users over the real number of spam ids

# 7-2 【SPAM相关】 如何test effectiveness of spam filter

1) 假如 FB 要建一个ML-based spam filter，用什么 measure 判断有效性?
    - Total number of spams reported by the user after the spams have been filtered (# of reported spam) or the report rate of spam
    - Engagement metrics: User engagement (# of posts, # of replies, like, share); Duration & frequency; Retention rate; # of outbound click (点击spam跳出FB网页)
2) 从 user experience 的角度， 什么 measure 说明有效? User engagement.
3) 如何设计实验来说明? User engagement change before & after launching the filter, t-test
4) how do you form your control/experiment group样本是random挑选么?

我回答应该从至少report过一次的人里挑，因为一般人可能不举报，说明不敏感。

5) 如果还没有 launch，如何看效果?

AB test. Choose two groups of users: (1) Do not report spams often (2) Report spam often. Do t-test on user engagement measures difference. There should be no interaction between train & test; similar distribution on train & test for spam expose, user engagement, behavior (duration, frequency, # of friends, # close friends/total friends). It's better to avoid network effect, clustering randomization sampling.

6) 如果发现使用ML-based spam filter 之后， user time spent下降，是为什么?

Time spent may decrease because they spent less time on spam contents that are already filtered. Frequency & engagement may increase because every time the user opens Facebook can watch what they interested, so they may open frequently to check the news feed. Retention rate may increase and activities may increase.

7) 新的spam model之后revenue下降了。面试官确定了首先这个model不会touch到ads，就是说ads不会被filter out。并且DAU/WAU/MAU和time spend没有变化，也就是说user方面没有变化。那么可能的原因是什么。
   - When spam decrease, then users are more likely to see the content they interested, then screen scroll length decrease, so the user is less likely to see the ads. 采用新的model之后用户可能会花更多时间去看video之类的，那么用在ads上的时间就变少了。
   - 我问面试官revenue主要来自什么，面试官说是click ads。我说那么ads click的revenue主要可以break down成#user x CTR x price. 这个情况下只可能有变化的是CTR，也就是说因为用来新的model以后，这个平台的整体content质量更高了，那么user就更喜欢花更多时间去explore这些content，那么点击广告的时间就相对来说变少了，revenue也下降了。
   - CTR: Maybe ad click-through rate per user changes, because of the change in user distribution. Users who used to engage less due to spams tend to engage more after the spam filtered by ML-model, but these users are less likely to click on ads.

8) spam降了发现engagement也下来了怎么回事?

9) 一个升一个降如何衡量效果

Short term vs. long term: how much does revenue drops? User experience vs. revenue. Short term revenue drop vs. long term brand perception and long term revenue gain.

10) 基于实验的结果，要不要launch这个model/这套 算法上线后，Spam Rate下降, ads revenue 也下降，我们如何决策是否继续这套方案? based on your hypothesis, how would you try to verify it? what are the metrics you would like to look into?

Short-term vs long-term: how much revenue drops, user experience vs. revenue, for long-term,
brand perception and revenue gain

- 要考虑文字本身的信息，可用sentiment classifier找推销 / 诱导语气很强的，output 作 为 feature
- 考虑 post metadata: 发表时间，是否反复重复，有没有带有链接outbound link， 特别是去不安全网站的
- class imbalance问题： false positive比false negative严重，把正常用户的帖子删 了， 后果比漏掉一个spam要严 重。 解决方法：
  i. 直 接 修 改 cost function
  ii. 用 bayes rule 预 测 cost 数 学 期 望最 低 的 class 。(别 用 upsampling, downsampling 就 行)
- 怎么衡量是否work：我没说AB testing, 最直接的显然是看reported spam是不是变少了, 然后看predicted spams和人工识别出的spam的交集。 这个算法不可能完全自动化，只要不是100%铁定的 spam就一定要有content reviewer把关的。衡量这个model的价值应该从content reviewer减负的角度出发, 如果不remove spam post, 只是降低它们在news feed的排位,

11) 如果不remove spam post, 只是降低它们在news feed的排位，怎么测量这个算法成不成功，具体用什么metrics, 怎么设计test，sample要怎么选。
12) fb目前identify spam是用如下方法：有人report一个post是spam之后，fb内部有人会review这个report，然后判断是不是真的是spam，是的话就删除。
    i. 请问现在这种方法有什么缺点: low efficiency, inaccuracy
    ii. 如果我们用一种算法来filter spam，你会用什么metric来衡量这个算法?
    iii. 问了问ab testing的步骤，confidence interval是什么

# 8-1 fb的friend suggestions 先描述了一下，问我觉得有什么意义，为什么要做这个。说了一下fb在获取new user上的挑战（因为说到这一切都是为了用户群更大），其中提到了怎么看真实用户群体，

1) 如果一家人就一个手机，该怎么判断这些account不属于同一个人。用什么metric看
On what time of the day the users use it; what kind of activity the users do; who they interact with (other users); the content and type of posts; similar id/name; shared friends
2) 问我认为它的goal是什么，怎么measure它的performance
Goal: increase level of engagement and retention
Metrics: # of friends; level of interaction (comments, like, messages); time spend on the product in general; # of referrals; retention rate
3) 如何判断这个algorithm是成功的，哪些人会用这个功能，这个功能和别的加好友的方法比有什么pros and cons
    - Increase rate of friends; complaint rates; rate of dismiss friend relationship
    - People with higher demands of social network; business account
    - Pro: remind people of friends they may ignore; reduce the # of search on finding friends; increase potential network
    - Con: can be disturbing and impair user experience
4) 怎么判断这是不是额外需要的
    - Did we receive any complaints or some findings in experience research?
    - Did the competing products provides similar features?
    - ROI analysis and simulations on the feature (small sample experiment & DID)
    - Small-scale AB test
5) 如果一个metric是多少，那么这个数字有什么意义
    - Significant (statistical & practical)
    - Cost and benefit (ROI)
    - Sanity check (sizing & AA test) + invariance + sign check
6) 如果一个metric降低，那么如何判断这个是有意义的
    - Significant (statistical & practical)
    - Outliers/spam
    - AA test
    - Seasonal change
    - Other experiment or marketing/PR events
    - Slice users into different categories
7) 哪个metric是最重要的: engagement
8) 朋友太多有什么不好的地方; 如果想要减少朋友数应该怎么做? 怎么找到推荐的人选? 怎么测试
    - 会使得newsfeed混乱，user experience 变差
    - 可以测试一下推荐取消朋友关系的功能
    - 可以建 model， 追问数据哪里来，可以看历史数据
    - a/b 测试

# 8-2 找好朋友题

1) 用一个metric来找一个人的最好的朋友。

- metrics:
  - # of like/comments/message with a specific friend / total # of like/comments/message
  - Sum of # of comments + like + message (with weights if necessary)
- filter: 成为好友的时间，去除共同的last name的

# 8-3父母加入fb的影响

1) 如何分析这个影响？看什么metric？怎么建模?

- 这个用户的activity可能会减少。 可以看sessions per week （一个月内用几次）
建模：quasi-regression (difference in difference) 父母是为加入作为一个0 or 1的 dummy variable，fit 父母加入前后的用户的 sessions per week

2) 如何找control group，哪些些feature需要match?

control group是父母没有加入的用户。要control for confounders。 比如 previous # sessions per week，父母与user的互动， user 的demographic etc

3) 如何判断哪些人有父母加入？只用fb提供的label的不好的地方。如果建模判断，错了type 1 error或者type 2 error 有什么影响?

可以看Facebook的relationship有没有显示family。
只看fb提供的label的话，说明用户不介意自己的父母加自己的Facebook有selection bias。
type 1: parents make a different but actually not (take actions that may not influence anything, so not a big deal)
type 2: parent does not influence users but actually does (do not take action but things go worse, can be a bad thing)

4) 父母都在不同时间加入的fb，对分析有什么影响？怎么办

比较困难set一个pre and post date。 如果父母加入的时间比较相近，那就把加入的min date之前的作为pre，max date之后的作为post来分析。

5) 要观察多久data看影响？一周？还是别的？why? 一周可能太短了。

6) 怎么看待分析的结果，如何break down结果？哪个因素最有可能有影响？如果找出来了，如何对产品提建议

可能可以breakdown by parent user interactions or age (other demograohic features)。可能有些用户不介意自己的父母看自己的fb。如果有些用户segment因为父母的加入减少来activity， 我们可以提供一些功能不让父母看某些信息（分组功能）。

7) 父母来了，子女走了，如何在父母和子女之间取得平衡?

从用户活跃的角度： 父母在fb上contribute 的 sessions vs 子女离开 fb 失去的sessions
从Revenue的角度： 父母在fb上的ad clicks vs 子女离开失去的 ad clicks

# 8-4 Closefriend notification

如果给用户的发text notification 告诉他们close friends 的Update，怎么评价是否需要加这个feature？各种detail和follow-up

- What is the goal? Engagement
- Metrics:
    1) view close-friend contents / all views
    2) engagement: time spent, sessions length/interval
    3) platform: retention rate, DAU, MAU, revenue
- AB Test:
    1) random select users into the experiment (sample size by alpha, beta, variance and effect size)
    2) push notification to the treatment group for a week (or longer duration by sample size and population size)
    3) Z-test on metrics
- Conclusion:
    1) statistical and practical significance
    2) sanity check (sample size), AA Test, Sign check

# 8-5 Unfriend

1) Pros and Cons
    - Pro
        - less worry about privacy (do not want some people to see my post) and may create more posts/contents, so engagement increase
        - find more good and meaningful contents, engagement increase
    - Con
        - spend less time on FB and revenue decrease
        - feel a weaker social network on FB (impression on less friends are using FB) so retention rate may decrease

2) 如何判断谁需要unfriend功能
    - session length/interval
    - interactions with friends
    - time spent
    - # of posts/contents
    - retention rate
3) 我们现在推出了这个feature, 怎么确认work？怎么判断结果到底有没有difference?
    - before launch: AB Test
    - after launch: observational studies (DID)
4) how to do z-test and what is p-value
    - find the metrics: CTR, engagement, retention rate
    - Z test score (diff-0/SE)

# 9. Instagram video call

1) We only add this feature in Spain for now, why we do that?

- Find one or two countries to test the new feature before completely launching to the whole platform
- Competitors in Spain lauched similar feature and had a success
- Did we receive any complaints or suggestions from the users in some other research (e.g. user experience research)?
- Is there any big change in user's behaviors or the engagement in Spain? If so, we may want to design some new features to increase the level of activity and engagement
- How large is the user size in Spain? It could be a good marketplace to test whether we want to launch a new feature to the whole platform; because we can diverse users (big size of users, as well as different categories of users), similar culture and social background to other countries (EU and US), large size of Spanish-speaking users

2) Imagine the head of Instagram walks in to your office, how do you tell him the effect of adding video chat feature to people's time spent on instagram?典型的AB Test问题
   - Do the AB test and prepare my report/slide, including:
     i. Background of the research (why we need it and what is the feature)
     ii. Major findings in the beginning
     iii. How we choose the sample, when did we do the experiment
     iv. What metrics we use
     v. Methods and assumptions
     vi. Findings and visualization
     vii. Compare the findings to relevant research (qualitative/quantitative)
3) How do you choose the control group? 要怎么做AB test （选择两个marketplace进行测试）然后还有一些dive deep的问题 比如如何选择两个接近的marketplace 要看那些metrics 如果短期没有什么improvement 要不要launch。
   - Before lacunching: 在Spain里分control vs test
   - After launching: Test group就是Spain整个country
     i. developed country
     ii. 地理位置要和spain挨的近
     iii. user behaviors要和Spain similar
     iv. demographic distribution也要相似啊
4) Suppose the result of the AB testing shows only 1% difference between control and test group, and we expect more, what is the reason?
   - Data accuracy & any spam or outliers in the data?
   - Statistical & practical significance (how do we set up the practical significance?)
   - AA test, sanity check, sign check (are we assuming our pick of control group is correct)
   - break down the population,see the performance of each age group, sex, etc.
   - seasonality
5) 假如新产品上线，例如算出某个metric 是20%，怎么知道是好还是不好
6) 做ab testing，应该在全世界抽2% 做random sample，还是只在一个国家测试，但是用这个国家所有用户做sample。
   - if the final goal is to launch worldwide, then it's better to use the random sample
   - if we want to know about the regional difference, we can sample by continent/region/language

# 10. Facebook有一个类似于slack的产品

1) we recently identify that the average number of onboarded are decreasing recently, what are the reasons you can think of?
    - Data accuracy; outlier and spam
    - 既然是average， 所以我的第一个想法是具体是怎么average出来，对于不同region和不同country， 具体是增加还是减少可能会不同，所以应该看看具体的region和country
    - 很有可能出了别的新的竞争产品，大家更喜欢这个竞争产品，所以自然用Facebook产品的人少了
    - Seasonal change有一个timeframe,那么很有可能某些公司已经有大部分人早就已经开始用了，没有用的人本来就很少，那么自然最近onboarded的人变少了因为整体的基数存量已经很少了。
    - Slice users into different categories
2) we are recently thinking about publishing a DO NOT DISTURB function, what are you going to do to determine if you are going to publish this feature.
    - AB test
    - Metrics:
        - CTR of the feature
        - level of engagement: time spent, session length/interval/#
        - platform-level: retention rate, DAU, MAU

11. We have a team that works on internal troubleshooting cases from clients. Assume that the reps prioritize cases based on a first come first served model and that you would now be responsible for prioritizing cases. How would you build a ranking system to prioritize cases?

# Common Questions

1. **增加 A NEW feature to a marketplace**

1) 其他marketplace 是否有launch这个feature， 如果有看看已经launch同类feature marketplace 表现
2) 如果没有，看看这个market pre-existing interest, 可以从内部用户浏览和点赞数据 也可以用外部已有的产品已有feature作分析，看看是否有demand
3) 如果有demand， 看看这个demand 是否能translate to engagement metrics 进而translate to ad review or MAU. 如果demand size能significantly提升，则考虑进一步做survey or A/B testing for the prototype.

**2. 如果一个metric突然drop 了, 你会从什么地方查找原因，用什么数据**

1) Clarify: what specifically dropped (metric used), by how much (practically significant/statistically significant)? -- if not significant then no need to go on
2) Data error & outliers (filter spam)
3) Then High Level:

a. Is it one-time or progressively? (One-time significant drop could be tech issue) -- One time is highly likely a tech issue. Seasonal is also ok
b. Does the drop happen in other features? -- If also other features then we have a bigger problem
c. Drop also happen in competitor products? Maybe competitor launched something new? -- if yes then may be a cross platform industry issue

4) Then Deep Dive (if anything changes in one of the segments; or maybe nothing changed but the distribution changed):
a. New user vs old user (Cohort)
b. Language
c. Country
d. Platform

**3. counter-metric/ DAU变多，但用户活跃度变少 / 以社交网络为例子，他们改变friends recommendation 算法，希望用户可以通过这个功能多加好友，如何衡量这个feature是不是成功呢。**

1) 我们先明确这个功能的目标，是希望多加好友，加了好友以后，希望在平台上有更多的交涉。那么这个时候的metrics可以应用funnel的思路来定义：好友增加率（friends request/accept rate)，月活量（monthly active user)，参与度(engagement），这些方面来定义metrics，尽量想到至少3个。
2) 然后做ab test，比较两个group的差别。
3) 这个时候面试官就会有一些follow up的问题，比如我们发现monthly active user(mau)增加了5%， 但是engagement 减少了3%，那么这个产品是好是坏呢。这个时候engagement就是我们的counter metrics，因为我们不仅希望加了好友以后，每个人的connection增加了，我们更看重是不是有意义的connection，所以engagement也是一个很重要的指标。一般这种题目可以从短期vs长期效应来考虑。比如我们的mau增加了，虽然短期的engagement减少了，但是长期看的话，由于每个人的connection增加了，由于network effect，大家相互影响，engagement长期看也许会增加的，这个时候就要去衡量long term effect了。

**4. 分析android和iphone用户的可能的差异以及原因**

1） Develop
2） Product analysis
3） UI design – logic of operation
- Slide: iOS goes to the next/last page vs Android switches tabs
- Tabs/features inside

4） Users:

i. Developing for Android makes sense if your target audience is significantly focused on Android, as well as when you're focusing on customizing Android user experience and adding to the personalization power that Android has

ii. iOS first makes sense when your goals align with the consumer spend, high in-app engagement and loyalty that makes iPhone users valuable; Apple has also reduced approval times and the time to market is working to your benefit

- Android has the largest global share in lower income areas and developing nations. It holds an advantage over Apple in emerging markets such as Asia and Africa. Apple, however, dominates the profit share despite Google’s global

dominance of market share because the average iOS user is more active that the average Android user.

- iOS users typically have higher income, higher education levels, more engagement, and spend more per app. Of course, that doesn't mean that those who have those same characteristics won't own an Android device. Instead, this data is simply indicative of the general Android population. Men are slightly more likely to be iOS users than women. Android seems to be the most common platform among all age groups, but its edge over iOS was a bit smaller in the 65+ age bracket.
- Android users are less extroverted than iPhone users, and they are perceived to have greater levels of honesty and humility. Android users are also more likely to prefer saving their money and to say they tend to follow, while their iOS counterparts prefer spending their money and they're more likely to say they tend to lead.
- iOS users are more likely to be early adopters and to have first used the internet before 1992. They also seem to be loyal to Apple, as they are more than 100% more likely to own a Mac computer compared to Android users. On the other side of the spectrum, Android users seem to be late adopters and they are less likely to backup their computer. They prefer a full-featured device at the expense of its appearance, and they are more likely to use Yahoo Mail as opposed to owning an email domain associated with work or their website.
- About 3.5% of Android users open push notifications, while just under 1.8% iOS users open them. One of the reasons for this discrepancy may be the fact that push notifications on Android stay visible on the lock screen until the user actions on them. On an iPhone, they disappear after the first screen unlock.
- On average, iPhone users engage with their smartphone apps for nine more hours in a given month than Android users. iPhone owners are sometimes described as smartphone "power users" and tend to engage with more content on average. On another hand, the Android platform has a greater number of media users in each category.
- The median iPhone app user earns $85,000 per year, which is 40% more than the median Android phone user with an annual income of $61,000. And even though Android has far more downloads than iOS, iPhone users spend twice as much as their Android counterparts. Also, in Q1 of 2017, iOS spending jumped 45% year-on-year, while Android's grew 40%.
- The average in-app shopping check is four times higher for an iOS user! If you'd want to develop a mobile shopping app, iOS development would make all the sense. However, Android users love digital utility apps like launchers, anti-virus apps and performance boosters, and they are ready to pay for those and spend on them 5 times more than iOS users. iPhone owners are also more likely to make purchases on their phones on a regular basis. These are important considerations for both retail app developers and those seeking to monetize via paid apps or in-app purchase. Mobile ads are the main source of revenue generation in Android apps.
- iOS users are more loyal and have more spending power. As I touched on briefly earlier, Android users love the openness of the platform and they like to customize their device, while iOS users like to keep their devices straightforward and spend more time on trying out various apps. It seems that iPhone owners tend to think very highly of their devices, and are likely to remain iPhone users over time as a result of that. And while the percentage of highly satisfied Android

owners is fairly high – 48% – it is quite below the 62% of iPhone owners that feel the same about their smartphone.

# Facebook 面试

Monday, January 13, 2020 8:47 PM

假设你的目标就是analytics方向，你要准备的顺序或者花的时间应该是这样：sql/r/python+case study > experiment design > machine learning。所以重点就是sql+case study。我下面就把这三个主题应该如何准备给大家说一下，重点还是放在case study上面

- SQL/PYTHON/R: 这里主要就是考察data munipulation，你能不能把复杂的逻辑转换到代码执行上面。
- SQL:sql我暂时没有看到特别好的资料，觉得最有用的还是地里的面经。如果真的一点都不懂的话，可以看看w3 school上面的题目，再把leetcode上面database的题目练习一下就行。最后面试前还是得刷面经，重点是逻辑一定要清楚,边缘case考虑到。我遇到的sql常考的有：各种join, union/except/intersect, 各种aggregate function（sum, ncount, average), case when, windows function（lag, lead, rank)。把我说的这几个都懂明白了怎么用，应该应付一般的sql没啥问题。不过sql很多时候都是逻辑比较复杂，真正写起来并不复杂，这种只能靠多见题目了，没有啥捷径
- Python/R: 我主要用python，考的题目无外乎就是把sql题目再用python写一遍，基本都是用pandas来写。如果对pandas不太熟悉，可以看看kaggle上面别人用python写的代码，主要看他们如何做data munipulation那块。另外对基本的数据结构也要了解，比如什么是list, tuple, set, dictionary。各大网站都有很多python的公开课，有时间就自己去上上吧，毕竟打好基础挺重要的。实在没时间就把kaggle别人写的代码弄明白了就够用了

- Case study: 重中之重！这块有点像咨询公司的case interview，也是考察能不能把问题break down，能不能发散把各种情况想的比较全面。不同的是，我们ds面case interview，更关注从数据的角度去解决问题，从用户使用产品的life cycle出发去定义metrics。但是我还是建议去看一下咨询公司面试题目，学习一下他们答题的stucture，如何跟面试官保持一个很愉快 conversation。这里推荐Victor Cheng的case study interview, 可以从喜马拉雅fm里面听，很适合上下班的时候听：http://www.ximalaya.com/5269453/album/6414597?feed=reset。 下面我来总结一下典型的case study常见题目以及答题思路。这类问题并没有一个统一的答案，只要是合理的就可以。

- Diagnostic problem: 这类问题通常是发现某个kpi某天突然下降了，该咋办？这类问题就是考察你能不能把一个business问题一步一步的break down，找到问题所在。一般的思路都是先从数据搜集的是否正确，是否这个kpi有seasonality，这个kpi有哪些不同的segment，等等这几点出发。举个最简单的例子：某社交网络发现用户使用likes的功能今天突然下降了10%，你觉得是为啥？首先你可以问面试官，咱们的数据搜集的正确与否，这个功能是不是有seasonality, 是不是今天有个speicial events发生（比如自然灾害，大家都断网了）。切记要keep this as a conversation，不要你一个劲儿的在那里说说说。根据面试官给你的引导，下一步就是如何break down, 你可以先说，likes下降了是指哪个部分下降了，有friends post, pages, or events。你还可以从另外一个角度break down，

比如是哪个地区的likes下降了。记住，understand the context is very important!。不要先急着答题，要先把数据的背景了解清楚。基本你把背景了解清楚了，也就找到了问题所在了。

- How to measure the success of a new product/feature: 这类问题一般就是我想改变或者增加一个产品的功能，如何衡量是不是成功了。这种问题的思路一般都是先搞清楚产品的目标是什么，你可以问面试官：what is the goal of this product? 从目标出发，先定义衡量的metrics，然后就是做实验了，根据实验结果来判断是否成功。在定义metrics的时候要想的全，哪些metrics对你的目标是最重要的，另外也不要忘了定义counter metrics，我在下面会用一个例子来解释。定义metrics注意一点就是不要仍给面试官一堆list，想3个最重要的就可以了。做完实验后面试官会有一些follow up问题，比如metric a 增加，metric b减少；或者metric a增加了2%，那么这些情况下是否应该launch这个产品呢。我下面用一个例子来说明这类问题的思路。还是以社交网络为例子，他们改变friends recommendation 算法，希望用户可以通过这个功能多加好友，如何衡量这个feature是不是成功呢。我们先明确这个功能的目标，是希望多加好友，加了好友以后，希望在平台上有更多的交涉。那么这个时候的metrics可以应用funnel的思路来定义：好友增加率（friends request/accept rate)，月活量（monthly active user)，参与度(engagement），这些方面来定义metrics，尽量想到至少3个。然后做ab test，比较两个group的差别。这个时候面试官就会有一些follow up的问题，比如我们发现monthly active user(mau)增加了5%， 但是engagement 减少了3%，那么这个产品是好是坏呢。这个时候engagement就是我们的counter metrics，因为我们不仅希望加了好友以后，每个人的connection增加了，我们更看重是不是有意义的connection, 所以engagement也是一个很重要的指标。一般这种题目可以从短期vs长期效应来考虑。比如我们的mau增加了，虽然短期的engagement减少了，但是长期看的话，由于每个人的connection增加了，由于network effect，大家相互影响，engagement长期看也许会增加的，这个时候就要去衡量long term effect了。然后面试官可能会接着问，假设我们的metrics都正向增加了，比如2%，那么如何判定这个2% 是个好的增长，可以launch to everyone? 这个时候可以往量化方面想，比如2%的增加对应了多少popualtion，这个2%的增加带来的潜在revenue是多少，如果是好几百万的话，那即便2%，也是个很好的提升！

- How to identify opportunity: 这类问题一般都是在做ab test之前，我如何去说服别人我的假设是合理的，我的假设值得去做一些test来衡量impact。这种问题很重要的一点就是identify opportunity sizing。也就是你的假设会影响多少人，如果只影响到5%的人，可能你带来的影响不会很大，但是如果超过20%的话，这个时候你的影响就大了。比如某电商想加个产品线，这个时候你如何去说服你的领导这个产品线值得去做一些实验呢。首先你要告诉你的领导，我加了这个产品线以后，会影响多少人，会带来潜在的多少收入。接下来就是如何去做这个产品线，比如我们应该target目标群是啥，应该以什么样的方式让目标群更好的了解我们的新产品线。再接下来就是如何做实验了，这样就回到了上面的how to measure the success的问题。从这里大家也可以看出，data science其实是一环套一环，有一个完整的周期的，我推荐的那篇airbnb的blog也说明了这一点。
- 实验设计：我自己的统计基础比较弱，这里就不班门弄斧了，就简单给大家说一说常见问题以及准备资料。

- 准备资料：在前一篇帖子里提到过：AB testing by Google from udacity

- 常见问题：如何randommize sample, 如何决定sample size, 决定sample size的几个因素对sample size如何影响，如何决定test跑多久，什么是p value, confidence interval, type 1, type 2 error，熟悉t test, z test公式跟原理。要注意不能只把定义背下来，要真正理解了，并且可以给non technical的人解释清楚。可以参考penn stats的假设检验章节：https://onlinecourses.science.psu.edu/stat414/node/290

- Machine learning analytics职位对ml要求很低，很多公司基本不问。问的话也都是基于你简历里面提到的项目，一般都是公司过去或者正在解决的real life problem，这种open-ended的问题千万不要想去找正确答案，因为并没有(摊手)，interviewer考察的是面试者拿到一个问题后how to approach the problem in a structured way，所以逻辑性非常重要，答案是否和面试官的想法一致是其次，言之有物即可，当然如果你逻辑清奇，偏离了比较主流的方向，面试官也许会觉得很新鲜顺着你的思路走，或者把你拉回来，都不用太担心。经验少的同学一般从套用先人们总结的framework做起，逐渐慢慢内化为自己的custom issue tree用来解决不同的问题，回答问题时一般可以考虑下面这几步：

  **1. Use Framework to structure your problem!**

  Ask clarifying question: what it does? how it's used? who it's for?
  拿到case一定不要上来就分析！你听明白情境和问题了吗？可以问一些边界问题缩小范围吗？(US only? narrow down to core product/business?) 该产品的功能你用过吗了解吗？不懂的问题一定要在这个时候问，如果分析到一半你问interviewer你说的这个功能怎么work的是这样吗是那样吗？你觉得对方会怎么想？有些产品就算你在开始说不太熟悉然后跟面试官当场确认主要功能也是totally fine的，如果去面FB然后说没听过newsfeed的另当别论。

  2) Verify high-level business objective & Define metrics that align with the objective
  非常重要！是要increase revenue, engagement, user base? Acquire new users or retain existing users?
  确定了对方想要达到什么目的才能制定相应的KPI
  对于需要时间衡量的long-term metric，要找到它的short-term proxy：if you can move that, you would expect to improve the long-term one.

  3) Hypothesis + Issue tree analysis
  Form a hypothesis and break down the problem using a framework or custom issue tree.
  这个就需要通过不断练习总结经验了，没什么捷径。
  举个栗子：比如问为什么今天DAU突然下降了，可以从这几个角度分析

  External factor: Seasonality(compare to same period last year)?; Industry trend? New competitor product launch? Special event, natural disaster?

  Internal factor: Data error, system issue? New feature release? Marketing event?
  Always break it down and decompose the metric: e.g. DAU = existing users + new users +resurrected users - churned users

  Then further segment by user demographic & behavior: location, device, platform, language, source, browser, etc. 可以exploratory analysis 也可以用点简单的ML来build a simple decision tree to identify

important features.

还有一些很好用的frameworks，比如：

- Awareness, Acquisition, Activation, Engagment, Conversion, Retention, Revenue, Referral：AARRR模型lz自己的拓展，根据需要增减
- Sales Funnel: total #visits to home page -> product list page CTR -> product detail page CTR -> % add to cart -> % check out
- Cohort analysis for user retention, LTV = Revenue/Churn rate
- How to increase market share? Acquire new users in existing market or expand to new market; increase user retention; develop new use case (eg. UberEats), categories; new technology(e.g. autonomous driving), just to name a few
- Opportunity sizing to evaluate a new opportunity, or proiritize features
- Pricing: cost + reasonable margin as floor, ultimate underlying benefit created by the product as customer maximum willingness to pay (ceiling), competitor pricing is somewhere in between
- Porter's five forces

4) Recommended Actions!!!!
分析data是为了什么？当然是给business partners提供actionable insights呀！所以分析的最后一定要落实到可执行的建议上！这才是你的价值！
常见的建议一般落实到product+marketing上
比如对找到的outperforming sector，可以尝试从产品方面制定loyalty program，或marketing来up-sell cross-sell，or acquire more of those users via marketing campaign
对于underperforming sector，看看如何改进产品remove friction，或者从marketing角度send reminder, incentive(coupon), retargeting email, etc

5) 比如准备采用4里面的一个产品改进建议，在决定launch之前要做什么呢？那就是A/B testing的主场啦
udacity的A/B testing by Google强烈推荐，一定要把知识点吃透(eg. how long do we run the experiment?)
面试过程中一般marketplace公司比较爱考selection bias, simpson's paradox, learning effect，又或者不能A/B testing的时候怎么办？

Udalicy A/B testing course
https://classroom.udacity.com/courses/ud257

Stellar Peers: 很多产品题的sample Q&A，可以重点看看product launch，strategy，marketing，pricing一类的文章，非常有助于开拓思路和structured thinking，然后照葫芦画瓢哈哈
https://medium.com/stellarpeers

Victor Cheng：Look Over My Shoulder(LOMS)：Consulting届的面试宝典，都是实战case study，其实这个跟tech公司面product题还是挺不一样的，毕竟面试官不是consulting出身的话他在你身上找的点还是跟面consulting不一样的，但是建议看看Victor讲framework的免费课程，非常有助于case入门，你会发现，原来都是套路呀！
https://www.caseinterview.com/case-interview-frameworks
LOMS audio：https://www.ximalaya.com/shangye/6414597/

Given the table about posts
date | user_id | post_id | content_type | extra

- **content_type**: 'view', 'comment', 'photo', 'report'
- **extra**: post text, comment text, report type = {'SPAM', ...}

Q: What is the total number of posts posted yesterday for each type of report?

Q2: When a post is reported by the user as spam, the reviewers will manually review and remove them. Given the table about review_removal

date | reviewer_id | post_id

What percentage of the posts viewed by the users yesterday is reported as spam?

然后是产品题，有关 spam filter。
Q: 假如 FB 要建一个ML-based spam filter，用什么 measure 判断有效性?
A: Total number of spams reported by the user after the spams have been filtered.

Q: 从 user experience 的角度， 什么 measure 说明有效?
A: User engagement.

Q: 如何设计实验来说明?
A: User engagement change before & after launching the filter, t-test

Q: 如果还没有 launch，如何看效果?
A: Choose two groups of users: (1) Do not report spams ofter (2) Report spam often. Do t-test on user engagement measures difference.

Q: 如果发现使用ML-based spam filter 之后，广告点击率也下降了，是为什么?
A: Maybe ad click-through rate per user changes, because of the change in user distribution. Users who used to engage less due to spams tend to engage more after the spam filtered by ML-model, but these users are less likely to click on ads.

SQL:
-- Q1: how many posts were reported yesterday for each report Reason?
-- Table: user_actions

-- ds(date, String) | user_id | post_id | action ('view','like','reaction','comment','report','reshare') | extra (extra reason for the action, e.g. 'love','spam','nudity')
-- Q2: introduce a new table: reviewer_removals, please calculate what percent of daily content that users view on FB is actually spam?
--no need to consider if the removal happen at the same post date or not.
-- ds(date, String) | reviewer_id |post_id

Product:

- How would you test if this filter works?

- List some metrics
- If we experiment, how would you conduct it?

- A/B testing
- How to select a sample group

- Random to avoid bias
- How many people would you select for your sample group

- Use formula for n (minimum sample size)
- After getting results from A/B testing, what to do next?
  . check 1point3acres for more.

- T-test on metrics to see if there's a difference-baidu 1point3acres
- What's a t-test? What's t-score? What's P-value? Explain p-value to someone who doesn't know stats.
- Let's say the filter worked but revenue went down, what would be your hypothesis?

- If it's rev/user dropped, perhaps the number of users increased because of the effective spam filter, but they don't spend money.

- But aside from rev per user, total revenue also dropped. Perhaps user distribution changed, there're more users who don't spend money now.

- Perhaps the filter downgrades the spam posts and now all the spams are clustered at the same spot (e.g. the 20th post and forward). So users stop looking scrolling after the 20th. Check activated time to validate.
- Given revenue decrease, how would you make recommendations? (doesn't have to be yes or no answer)

- Short term vs. long term: how much does revenue drops? User experience vs. revenue. Short term revenue drop vs. long term brand perception and long term revenue gain.

- If user distribution changed: find cause and tackle that unimaged users

Q1: how many posts were reported yesterday for each report Reason?
Table: user_actions

ds(date, String) | user_id | post_id | action ('view','like','reaction','comment','report','reshare') | extra (extra reason for the action, e.g. 'love','spam','nudity')

Q2: introduce a new table: reviewer_removals, please calculate what percent of daily content that users view on FB is actually spam? no need to consider if the removal happen at the same post date or not

ds(date, String) | reviewer_id |post_id

Q1: what metrics to validate if the spam filter algorithm work or not?

Q2: how you validate if the spam filter algorithm work or not?

Q3: if the spam filter works but revenue drops, you know why?

Q4: how you will design the spam filter algorithm?

发个面经，回馈一下地里 面试官是个略有口音的三姐，因为极力想听清楚她在说啥导致我有些不必要地紧张……

第一道，分析案例题，问青少年父母加入后会不会造成青少年流量下降。主要考product sense，A/B Testing

追问，如何判断两个用户之间是否为父母/子女关系？

第二道，SQL题，数据格式：

表名：survey_log 列名：user_id, question_id, question_order, event = {saw, answered, skipped}, timestamp

具体题目：刚加入的用户会要求填一份调查问卷，但问卷里的问题也可以跳过，每道题只要被用户见到就会生成一条记录（event为saw），如果被回答或者被跳过会生成另一条数据（即每道题每个用户都会有两条记录），回答则event为answered，跳过即为skip。并且每道题出现在每个用户前的顺序有可能不同，所以有question_order。假设该表里已经存了1M用户的数据，在一个新用户进来时，如何安排题目尽可能多地得到新用户的答案，减少skip？

a. 先问how to measure health of facebook,我说可以分两个部分，一是website/sytem本身health，是不是function normally，另一部分是user engagement，比如新用户注册率，现有用户的interactions包括comment/share blablabla，然后DAU/MAU之类的，

b. 然后他直接打断说ok，现在就focus在comment上，如果要求你做一个dashboard specifically about comment，有哪些metrics你可以present。我先莫名想岔到之前面经的sql题了，跟他说可以show distribution of comments/users，然后他表现得不太满意，说ok，其他的呢，我就说可以present average comments per post/user，然后他继续问这个average comment per user是怎么计算的，denominator和numerator是啥，我回答numerator等于总评论数，denominator是distinct users who have left at least 1 comment per day。

c. 然后他说ok，现在把denominator换一换，改成DAU，我们发现这个总comment/DAU的比率跟一年前同一天相比，增加了50%,有哪些可能原因。我回答首先user mix可能有变化，可能今年新增的growth都是愿意发更多comment的用户，然后（随口乱说）content内容的变化吸引更多user leave

comment，第三可能有新feature刺激用户留言更多之类的。然后他follow up了一句什么我忘了，好像是如何判断是哪种？我就先说我们可以segmentation by users' language/platform/device type/browser/location看看有没有哪个subgroups特别突出，有的话可以dive deep。

因为书快100页了，我大概缩减成了几类问题，以下答案都是书里的浓缩，：

**1)经典题型 15% Drop in FB group usage:**

1) Clarify: what specifically dropped (metric used), by how much (practically significant/statistically significant)? -- if not significant then no need to go on
2) Then High Level:
a. Is it one-time or progressively? (One-time significant drop could be tech issue)
-- One time is highly likely a tech issue. Seasonal is also ok
b. Does the drop happen in other features?
-- If also other features then we have a bigger problem
c. Cannibalization
d. Drop also happen in competitor products? Maybe competitor launched something new?
-- if yes then may be a cross platform industry issue
3) Then Deep Dive (if anything changes in one of the segments; or maybe nothing changed but the distribution changed):
a. New user vs old user (Cohort)
b. Language
c. Country
d. Platform

**2)How to improve the product?**

The question is not asking you to be visionary. But to check if you can find things from datasets as a data scientist. Always try to incentivize "good" and dis-incentivize "bad".
1)Firstly, define the target. Say engagement (in order to move long-term retention and revenue)
2)Then choose metric used to evaluate engagement: i.e. the proportion of users who take at least one action per day interacting with the site.
3)Pick variables that would move the metric: use both user characteristics and user behavior
4)Use model (random forest is good here) to check the relationship between segment and engagement. Come up with several scenarios to explain and make suggestions based on the results (improve which segment)

**3)Fake/Fraud detection:**

Key with fraud is, not happening only once. People who commit fraud would like to repeat it if not being caught. all variables are really about something that should be unique but is not or extreme values. Hence two main ways to capture fraud:
1) Same device IP/Bank account/phone number as existing accounts;
2) Anomaly detection-find outliers (extremely low price)
Ø More specifically with market place posting, we can address the listing and seller. For listing, pictures cannot be stolen from elsewhere/descriptions cannot be copied/resolution should not be too low/price should not be two
Ø With fake profile (say fake school): using ML algorithm or anomaly detection to find outliers. For instance, you may include the percentage of connections went to the same school/interaction with people from the same school/acceptance rate for the same school request as variables. In order to minimize the fake profile, you may want to use 2-step verification for risky users (minimize false negative you may not apply this to all users).

**4)What features to add?**

Again, not tempted to be a visionary. Starting from the datasets. Look at current data and check where you want to incentivize people to do. Then simplify the procedures. You can also learn from customer needs through complaints or comments. Then A/B testing to see if it can satisfy your needs.
Eg: figure out a way for a user to finish things in one click/ check use case to find opportunities

**5)Should we introduce XXX feature?**

Layer of logic:
1) If add, what benefits will we get?
2) Do we have customer needs? (check from comments or user behavior)
3) A/B testing process

product里面有时会穿插一些AB testting的概念。这一部分我主要是看了udacity的课程 https://classroom.udacity.com/courses/ud257 看了三遍，大概对AB testing有了基本了解:

**A/B Testing Process**

1) Goal (increase revenue? Engagement? new user? old user?)
2) Metrics (invariant + evaluation)
a. Long-term use short-term proxy
b. Invariant is for sanity check
c. Think about how spam/bots would influence your metric
d. ! when choosing metrics, make sure the directional change of the metric is in line with your expectation and the change is unlikely due to bot behavior and it would not take too long to evaluate
3) Unit of diversion and unit of analysis?
a. If not the same, then need to use empirical variability
4) Size and Duration
a. Size is determined by alpha (significance level)/power(1-beta)/Variability of the metric
b. Using size and proportion of traffic applied, we can get duration (if greater than 14 days you're done; if less you may still need 14 days to know the weekly patterns)
c. ! Note that once size and duration are determined, you cannot stop halfway because the test result looks great and promising. The size is pre-determined in order to reach the alpha, beta, significance level needed. (same thing as a competition with 9 games, you cannot stop and announce the winner simply because one player wins 3 games in a row)
5) Analyze result
a. Sanity check (make sure test and control are comparable)
b. One metric is easier: construct confidence interval using diff + SE
c. Multiple metrics: false positive become more common as the number of metrics increases. use Bonferroni correction
6) Make suggestion
a. Do I have statistical/practical significance?
b. Do I understand the change? Who is going to be impacted?
c. Is it worth it? Cost vs benefit?

f

# 内推 2019 -12-23

**第一轮 HR邮件 2019-12-24**
问：毕业时间，最强编程语言，希望去哪个城市，目前VISA，什么时候来的美国，OPT还能用多久，是否需要H1B，有没有申绿卡，为什么FB。

**第二轮 HR pre-screening 2019/2020**
HR什么问题也没问，因为根本没有打电话，被鸽了。。。

**第二轮 HR pre-screening 2019/2020**
HR介绍工作内容，然后自我介绍。问了SQL基本问题：ORDER BY，LEFT JOIN，COUNT(DISTINCT)

**第三轮 技术店面 2018-11-26**
**寒暄：**对方介绍工作内容，然后我介绍一个最近做的项目，挑个跟未来FB工作相关的讲讲就好；
**SQL：**运气太好，非常简单，就是问广告conversion rate，两个table，一个ads，一个user，然后算算转化率啊之类的，很基本；我就用了一些LEFT JOIN，GROUP BY，ORDER BY，SUM，COUNT，IF之类的简单语法；
**Product：**怎么判断FB两个人是best friend。我先谈什么这个重要，然后怎么设计metrics，每个metrics可能有什么问题，最终怎么用ML做推测，并谈谈可能出现什么问题；

**Onsite邀请 2018-12-10**
时隔两周终于收到onsite邀请，本来已经放弃了都。等面完了就在补充里简单说说过程。

**补充内容 (2019-1-11 04:44):**
**Onsite 2019-01-08**
**1）Presentation：**30min (ppt) +15min (Q&A)，会录像，跟HR讲自己PhD做了些什么，要求至少提前一周把题目、摘要和自我介绍发过去。由于楼主PhD跟CS不沾边儿，便讲了讲大概idea；

**补充内容 (2019-1-11 04:50):**
**2）SQL：**30min，只给了一个table，3小问，不难，用到MAX，COUNT(IF)，GROUP BY，JOIN等基本语法。但中间问了其中一个我给的答案怎么提高efficiency，楼主说不会，他说没事儿；

**补充内容 (2019-1-11 04:53):**
**3）午饭：**45min，还呆在Lobby 6，大部分窗口排队人很多，遍找个人少的，但饭菜质量就selection bias了，而且人多到连坐的地方都没了，站着吃的。FB派了一个不面我的人陪我吃饭，聊天还是很轻松愉快的；

**补充内容 (2019-1-11 04:58):**

**4）Quant：**30min，很基本的概率题，用Bayes就够，算之前可以想想能不能直接得答案。然后会谈到FB产品中一些metric的统计分布，并让估计平均值，IQR之类的。楼主没用过FB，便胡乱猜了几个数，感觉凉了；

**补充内容 (2019-1-11 05:03):**
**5）Applied Data：**30min，如何利用data来分析产品中可能出现的问题。给了FB的两个明星产品让我做比较、分析以及建议，但是那两个产品楼主都没有亲自用过，所以只能回答得比较general，越来越凉了；

**补充内容 (2019-1-11 05:06):**
**6）Product：**30min，楼主认为最有意思的环节。要求从头到尾设计一个FB的新功能，这个倒和做科研比较像，遵循experiment design的顺序就好，中间可以有一些brain storm，但最好都围着data来谈；

**补充内容 (2019-1-11 05:12):**
以上就是楼主FB onsite的大致过程啦～由于签了NDA，所以只能说说考到了什么知识点，希望对大家有帮助！等结果出来了再来更新一下。祝大家FB DS面试顺利！

**补充内容 (2019-1-24 23:40):**
**Offer 2019-1-23**
HR直接打电话给的offer，具体内容请参见我的抖包袱。FB timeline到此结束了。最后祝大家FB面试顺利～

# Questions

## User 1

发个面经，回馈一下地里 面试官是个略有口音的三姐，因为极力想听清楚她在说啥导致我有些不必要地紧张……
第一道，分析案例题，问青少年父母加入后会不会造成青少年流量下降。主要考product sense，A/B Testing
追问，如何判断两个用户之间是否为父母/子女关系？

第二道，SQL题，数据格式：
表名：survey_log 列名：user_id, question_id, question_order, event = {saw, answered, skipped}, timestamp
具体题目：刚加入的用户会要求填一份调查问卷，但问卷里的问题也可以跳过，每道题只要被用户见到就会生成一条记录（event为saw），如果被回答或者被跳过会生成另一条数据（即每道题每个用户都会有两条记录），回答则event为answered，跳过即为skip。并且每道题出现在每个用户前的顺序有

可能不同，所以有question_order。假设该表里已经存了1M用户的数据，在一个新用户进来时，如何安排题目尽可能多地得到新用户的答案，减少skip？
求了个每道题的回答率，注意不能只求回答次数，要除以总的看见此题的次数
追问，即使按照回答率对题目进行排序，如果新来的用户已经skip掉了回答率最高和次高的题，如何动态调整题目顺序，获得此用户尽可能多的回答？
我在这题上挣扎了很久，思考了题目内容分类，题目之间的相似度分类，等等等等，最后发现其实是条件概率，要看用户之间的相似度，即已有数据中跳过了这两道题的用户回答率最高的是哪道……

然后并没有考说好的第三道概率统计题，直接聊了两句问了问题就byebye了

面完后跟进很快，HR第二天就通知可以安排onsite了，求bless！！

## User 2

电面是跟一个Sr. Data Scientist，大概45min，过程有三部分：
1. introduction，about 10 min，主要他介绍自己经历，然后大概问了我最近的工作经历，以及为什么跳槽；

2. technical，about 20 min，给了table format和两个问题，随便选语言来回答，我选的sql，问题跟http://www.1point3acres.com/bbs/thread-282664-1-1.html基本一样，表名：survey_log 列名：user_id, question_id, question_order, event = {imp, answered, skipped}, timestamp，第一问是找conversion rate最高的question，我在写完answer rate以后有个follow up问题，说如果imp太少的问题怎么办，我答的是加个threshold，舍弃那些出现少于多少次的问题（这里我随便写的10，他没有说什么，不知道对不对）。第二问是在用户已经回答了某一问题的情况下，如何安排下一问题使conversion rate最高，我这里就按地里讨论的一样说在已经回答了这一问题的用户中，选他们回答的其余问题里回答率最高的一个，但我太紧张了时间用太长，最后写的里面他说有bug，但我还没思考出来就说时间到了，我们进入产品题，我猜我因为这个多半就这么挂了。。

a. 先问how to measure health of facebook,我说可以分两个部分，一是website/sytem本身health，是不是function normally，另一部分是user engagement，比如新用户注册率，现有用户的interactions包括comment/share blablabla，然后DAU/MAU之类的，
b. 然后他直接打断说ok，现在就focus在comment上，如果要求你做一个dashboard specifically about comment，有哪些metrics你可以present。我先莫名想岔到之前面经的sql题了，跟他说可以show distribution of comments/users，然后他表现得不太满意，说ok，其他的呢，我就说可以present average comments per post/user，然后他继续问这个average comment per user是怎么计算的，denominator和numerator是啥，我回答numerator等于总评论数，denominator是distinct users who have left at least 1 comment per day。
c. 然后他说ok，现在把denominator换一换，改成DAU，我们发现这个总comment/DAU的比率跟一年前同一天相比，增加了50%,有哪些可能原因。我回答首先user mix可能有变化，可能今年新增的growth都是愿意发更多comment的用户，然后（随口乱说）content内容的变化吸引更多user leave comment，第三可能有新feature刺激用户留言更多之类的。然后他follow up了一句什么我忘了，好像

是如何判断是哪种？我就先说我们可以segmentation by users' language/platform/device type/browser/location看看有没有哪个subgroups特别突出，有的话可以dive deep。

## User 3

1. There is atable that tracks every time a user turns a feature on or off, with columnsuser_id, action ("on" or "off), date, and time.

1) How manyusers turned the feature on today?
USER_ID || ACTION||DATE||TIME

```
SELECTCOUNT(DISTINCT USER_ID)
FROMTABLE
WHEREDATE = CURDATE()
ANDACTION = 'on';
```

How manyusers have ever turned the feature on?

```
SELECTCOUNT(DISTINCT USER_ID)
FROMTABLE
WHEREACTION = 'on';
```

2) Create atable that tracks the user last status every day.

```
SELECT A.DATE, B.USER_ID, B.STATUS
(SELECT GENERATE_SERIES('2018-01-01'::DATE,'2018-09-01'::DATE, '1D')::DATE)
TABLEA(DATE)
LEFT JOIN
(SELECT * FROM
TABLE
QUALIFY ROW_NUMBER() OVER (PARTITION BY USER_ID, DATEORDER BY TIME DESC)
=1
) B
ON TABLEA.DATE >= B.DATE
QUALIFY ROW_NUMBER() OVER(PARTITION BY A.DATE, B.USER_IDORDER BY B.DATE
DESC) =1
```

3) In a tablethat tracks the status of every user every day, how would you add today's datato it?

```
/*ASSUMINGACCOUNTS KEY ARE UNIQUE IN TABLE_TODAY */
/*OTHERWISEWE CAN PICK THE LAST ACTION IN TODAY AS THE STATUS*/

SELECTA.*
```

```
FROMEVERYDAY_SATUS
UNION
(
SELECT
COALESCE(A.USER_ID,B.USER_ID) AS USER_ID
,CASEWHEN A.USER_ID IS NULL THEN B.STATUS
WHENA.USER_ID IS NOT NULL AND B.USER_ID IS NULL THEN A.ACTION
WHENA.USER_ID IS NOT NULL AND B.USER_ID IS NOT NULL THEN A.ACTION
END ASSTATUS
,CURDATE()AS DATE
FROM
TABLE_TODAYA
FULL OUTER JOIN
TABLE_EVERYDAYB
WHEREA.USER_ID = B.USER_ID
ANDB.DATE= CURDATE()-1

);

/*OR*/
INSERTINTO TABLE_EVERYDAY
SELECT
COALESCE(A.USER_ID,B.USER_ID) AS USER_ID
,CASEWHEN A.USER_ID IS NULL THEN B.STATUS
WHENA.USER_ID IS NOT NULL AND B.USER_ID IS NULL THEN A.ACTION
WHENA.USER_ID IS NOT NULL AND B.USER_ID IS NOT NULL THEN A.ACTION
END ASSTATUS
,CURDATE()AS DATE
FROM
TABLE_TODAYA
OUTERJOIN
TABLE_EVERYDAYB
WHEREA.USER_ID = B.USER_ID
ANDB.DATE= CURDATE()-1
```

5), 如何找出在一天之内始终保持feature on的人，(given status table and action table).

```
/*The only case is: YESTEDAY WAS ON AND THERE IS NO ACTION TODAY*/
SELECTDISTINCT A.USER_ID
FROMTABLE_EVERYDAY A
LEFTJOIN
TABLE_TODAYB
ONA.USER_ID = B.USER_ID
ANDA.DATE = CURDATE()-1
```

```
ANDA.STATUS = 'ON'
WHEREB.USER_ID IS NULL
```

2.

1) 有content_id, content_type (comment/ post),target_id。如果是comment，target_id就是post的content id，如果是post则target_id为NULL。求commentdistribution。

```
CONTENT_ID|| CONTENT_TYPE || TARGET_ID
123|| COMMENT || 100
100|| POST || NULL
-baidu 1point3acres
SELECTNBR_COMMENT
,COUNT(CONTENT_ID) AS NBR_POST
FROM
(SELECTA.CONTENT_ID
,COUNT(B.CONTENT_ID) AS NBR_COMMENT
FROMTABLE A
LEFTJOIN TABLE B
ONA.CONTENT_ID = B.TARGET_ID
ANDA.TARGET_ID IS NULL
ANDB.TARGET_ID IS NOT NULL
GROUPBY 1) C
GROUPBY 1
ORDERBY 1
```

2) 如果现在content_type变成post, video, photo,article，要求计算每一个content type的comment distribution。

```
SELECT
CONTENT_TYPE
,NBR_COMMENT
,COUNT(CONTENT_ID)AS NBR_CONTENTS
FROM

(SELECT
A.CONTENT_TYPE
,A.CONTENT_ID
,COUNT(B.CONTENT_ID) AS NBR_COMMENT
FROMTABLE A
LEFTJOIN TABLE B
ONA.CONTENT_ID = B.TARGET_ID
ANDA.TARGET_ID IS NULL
ANDB.TARGET_ID IS NOT NULL
```

GROUPBY 1,2) C
GROUPBY 1,2
ORDERBY 1,2

3) Story includes either photo or post . Generate a distribution for the#comments per story. Does this account for stories with 0 comments?
CONTENT_ID|| CONTENT_TYPE || TARGET_ID
123|| COMMENT || 100
100|| POST || NULL

```
SELECT NBR_COMMENT,
COUNT(*)AS NBR_CONTENT
FROM
(SELECTA.CONENT_ID
,/*TO COUNT STORY WITH 0 COMMENT*/
CASEWHEN B.COMMENT IS NULL THEN 0 ELSE COUNT(B.COMMENT) END AS
NBR_COMMENT
FROMTABLE A
LEFTJOIN TABLE B
ONA.CONTENT_ID =B.CONTENT_ID
ANDA.CONTENT_TYPE IN ('POST','PHOTO')
ANDB.CONTENT_TYPE IN ('COMMENT')
GROUPBY 1) C
GROUPBY 1
ORDERBY 1
```

4) 问怎么判断comment是一个有内容，有意义，conversational的comment
Definemeaningful and conversational.
General:

- Back and forthinteractive conversations, not just one user dominates the comments
- Positive signs such aslikes, shares, emojs, etc
- More interactions inbetween users because of the comments, both online and offline

Different types of posts may have different definitionsof nring meaningful:
Recommendations: good recommendations that is taken bythe poster;
Photos: Friends or families who were tagged leftcomments
Merchants: User who left comments as product feedbacks orresult clicking the ads, following the merchants, buying the product, etc.

3.1) 现在有一张表，有time, user_id, app,event（impression，click），每个用户在每个app上有一定几率弹出一个窗口填写信息，如果填写了event为click，如果只看见没填写为impression，没看见为空。求这个功能的clickthrough rate。

Time|| user_id || app || event

```
select
   sum(case when event = 'click' then 1 else 0end) clicks,
   count(*) as impressions,
    clicks/impressions CTR
from log_table
```

2) 如果CTR>100%是什么原因？
-data error (undercount the impressions) or delays from impressions
-some users click the link multiple times within 1 impression. Maybe they arehesitating whether to buy the product or not

3) 如果现在每一个impression可能对应多个click，如何从所有click记录里面选出正确的那个记录来计算？
Whatmetrics would you use in this situation, and what you should do next with thiserror?
Firstunderstand the 'error'. If it's data delay, then wait for the data topopulate.
If1 user clicks multiple times per impression, then cap it 100%, or dedupe atuser level.
Ifit's competitors who wrote a program to click the ads to mess up your metrics,then FB should stop it.
Butdespite the reason we can either dedupe at user level, or use click-to-actionrate

4) group 1 click rate: 10%, group 2: 15%, think aboutpossible differences,
Component:User, Ads (content), FB (how often)
User:
lessactive;
Doesn'tlike ads;
Activeuser vs. Prospect;
Mobilevs. Big browser;

Ads:less relevant (algorithm);
Biggervs. Smaller;

4. Givenan event-level table of interactions between pairs of users (note thatthere aren't duplicates in one day for one pair of users), for each possiblenumber of "people interacted with" find the count for that group in agiven day (i.e. 10 people interacted with only one person, 20 with 2, etc.).

U1||u2
Assuming that the table is consists of two columns, u1and u2:

```
SELECT
    countU,count(countU) ascountOfCount
FROM
    (SELECT
      uall,count(uall) as countU
    FROM
       (SELECT
        u1 AS uall
       FROM
        userPairsDF
       UNIONALL
       SELECT
        u2
       FROM
        userPairsDF)
    GROUP BY uall)
GROUP BY countU
```

given a tableof interaction between users (user_a | user_b | day), find number of users whohad more than 5 interactions yesterday (assume there is only one uniqueinteraction between a pair of users per day).

```
select usr , sum(cnt) as interactions
from
(
select user_1 as usr,
from f
where date = curdate()-1
union all
select user_2 as usr
from f
where date = curdate()-1
))
group by usr
having sum(cnt)>= 5
```

. From 1point 3acres bbs

5.

We arestudying ecommerce advertisers on FB over a certain time period (say a week).
The timeperiod does not matter for this problem. You are given 2 Tables:
adv_info:advertiser_id|| ad_id|| spend (primary: ad_id)
ad_info:ad_id|| user_id|| price (primary: ad_id, user_id)

adv_info:Contains information on advertisers.
advertiser_idis id of advertiser
ad_id is idof an ad being run by advertiser
spend isamount of money in $ that advertiser pays Facebook for ad-id to show it to FBusers.
price is how much the user_id spend through this ad.,assuming all prices > 0.
Questions- The fraction of advertisers has at least one conversion. - What metrics wouldyou show to advertisers.

Q1: What would the averageadvertiser spend on Facebook? Your query should return a single number.

```
SELECTAVG(TOT_SPEND) AS AVG_ADVSR_SPNT
FROM
(SELECTADVERTISER_ID
,SUM(SPEND)AS TOT_SPEND
FROMADV_INFO
GROUP BY1) A
```

追问：afterwe get this table, what will the distribution look like?
NOT SURE ABOUT THIS QUESTION

```
SELECT
A.ADVERTISER_ID
,A.SPENT
,COUNT(B.*) AS IMPRESSIONS
,COUNT(DISTINCT B.USER_ID) AS UNIQUE_USER
,SUM(B.PRICE) AS SALES
,A.SPENT/NULLIFZERO(IMPRESSION)AS COST_PER_IMP
,A.SPENT/NULLIFZERO(UNIQUE_USER)AS COST_PER_USER
,A.SPENT/Z
FROM
ADV_INFOA
INNERJOIN AD_INFO B
ONA.AD_ID = B.AD_ID
GROUPBY 1,2
```

Q3.Thefraction of advertisers has at least one conversion.

```
SELECTCOUNT(DISTINCT CASE WHEN B.PRICE>0 THEN A.ADVERTISER_ID
END)/COUNT (DISTINCTA.ADVERTISER_ID) AS PCTG_CONVERSION
FROM
ADV_INFOA
INNERJOIN AD_INFO B
ONA.AD_ID =B.AD_ID
```

## User 4

FACEBOOK. 1point3acres
Self-Introduction: 2-min

SQL: 30-min

Composer . 1point3acres
User_id | date | event {enter, post, cancel}

User
User_id | date | country| dau {1,0}
[/hide=100]

- Success post rate each day for the past 7 days?
- Average post by daily active users by country today?
- follow up question: why use right join not join, what's the different between two output? — some active user might not in composer post as we want to calculate avg, this matters.

Product: 25-min
[/hide = 100]. From 1point 3acres bbs

- Avg post nums 3.0 2018-05-01, 2.5 2018-06-01, what would you investigate further?
- Add composer tab, design metri

## User 5

上次写了第一轮电面，上周五刚飞去昂赛，身心俱疲，先写一下整体感受
可能是我本身没有做过互联网行业，所有的产品sense都来自于纸上谈兵，虽然准备了很久，但面完还是觉得心凉凉

先说下整体感受，特别是和我本人expectation差异比较大的地方：

- round 1：第一轮产品，主要聊选metric，顺带会问下AB TEST前期分组的注意事项。整个过程需要不断和他沟通，能看出面试官的反应
- round 2：统计/概率类的：第一题统计相关，刷到过真题，三小问，做完。第二题说launch一个feature后metric一个上升一个下降怎么回事（给了CI）。奇怪的是她让我从产品角度说，不是从数据角度说，感觉我最后也没说中她要的答案
- round 3: 另一个产品，更侧重对数据的处理，感觉他心里有个清晰的答案，你没按照那么走就一直打断你，follow up。这轮我有点被discourage，直接导致我午饭的时候一直情绪不佳，没有特别cheerful，对不起那么陪我聊天的开朗的小姐姐 QAQ
- round 4：SQL: 题目给完了，我又clarify了一些问题。前两问在射程之内，第三问很绕人。不得不吐槽一下大家都说SQL都是地里的题目，我的那个题逻辑比地里复杂很多呀～解决方法不太一样啊～童话里都是骗人的！第三问我用了个复杂的方法写完了，中间一度Interviwer都被我的方法绕进去了，抱头在那儿思考了好一会儿（苦笑）。但是貌似他还想和我讨论一个什么别的没时间了

总而言之感觉在网上看了再多面经，再刷题，昂赛还是很难把握局面
而且基本没有太问我做AB TESTING的内容啊～什么run time, sample size啥都没扯到，白白把ab test学那么多，欲哭无泪

感觉凉凉+身体被掏空

大家加油～～～～～

补充内容 (2018-6-15 04:31):
第一题：（我自己大概翻译的）assume now we don't have recommendation to users for what they post, if we want to add this feature, how would you evaluate?
一定要和面试官confirm一开始所有的细节，楼主事先告诉自己要communication但现场还是自说自话犯了低级错误。回家细想了下，这题还是蛮有意思的，user 采纳FB推荐post的内容多了，可能和整体user engagement 是个trade off。所以metric到现在我都没有结论到底应该选啥。我就不误人子弟了，点到为止

补充内容 (2018-6-15 04:35):. check 1point3acres for more.
第二轮： Q: We have two options for serving ads within Newsfeed: M1： out of every 25 stories, one will be an ad。 M2 ： every story has a 4% chance of being an ad
求两种情况下expectation & var

补充内容 (2018-6-15 04:36):
第三轮： if your supervisor wants to know what’s the total number of fake news today on FB, how will you get the number? 这轮我答的直接扑街...差的我的答案真的什么好借鉴的了

补充内容 (2018-6-15 04:42):
亮点在sql好么：每次User打开app叫做home load, 会看到10个stories. 其中有些可能是ads。再往下滑会继续看到stories。uder_id| timestamp| action (click/imp)| story_id. 亮点是我和他clarify才明白一个user_id最先在home load看到的那10个story的timestamp是一样的，接下来的这个人scroll down又看到很多story每个timestamp不一样，然后home load的十个应该是完全不同的，但是scroll down看到的可能已经出现在home load里面了。然后这个user一天内会多次打开app每次都有home load & scroll down…地里应该史无前例吧哈哈哈

补充内容 (2018-6-15 23:18):
概率题还有第三小问求 probability： (among 100 stories, see at least 8 ads)

补充内容 (2018-6-15 23:21):
楼主能刷题的都做出来了，连变态SQL都用了更变态的方法写出来了～（苦笑）～产品这个东西，我觉得我们这种跨行业跳槽的只能：多看，多交流，多面试，多fail

## User 6

刚结束了电面，印度姐姐人很好，会给各种提示，整体体验不错。
上来印度姐姐先自我介绍了一下，然后换我介绍目前的工作及内容。

以下问题欢迎大家讨论。
Product：
1. 一般说来一个user有很多friends是很好的，但是一个user有太多的friends也会有问题。你觉得会有什么问题？
答：太多newsfeed可能会导致user想看的东西看不到了，导致engagement降低

2. 如何来判断一个user的close friends
答：可以通过他们在一些特殊日子是不是会送礼物，照片里他们是不是会频繁出现被tag。

3. 如果有一个unfriend button， 如何向user推荐可以unfriend的人。
答： 可以给interaction（比如like，comment），互送gift， 照片里出现频率等factor加权重，然后算出一个score，根据score来排序。

还有一些不太回忆得起来，会有很多follow up questions，但是印度姐姐也会给挺多提示。我自己不清楚的时候也会restate question，保证不要答非所问。

table 1:
user1 | user2
123 456
456 123
123 789
789 123

table 2: sender | reciepient | action | date
123 456 create 2019-01-01
456 123 create 2019-01-01
123 789 create 2019-01-01

问每一对friends的interaction是多少？（一个create就是一个interaction）

## User 7

有一个表格叫DAU有ds, user_id。
请给出growth_accouting表格有ds, user_id, state.
state有如下几种(new, churned, resurrected, retained)

我感觉这个题目还是比较难的sql题目。很容易弄错。

## User 8

面试分三部分：
1. interviewer先介绍自己组的情况，然后大概了解了一下我的工作经历
2. analytical part（只给我面了一道题。。。没有产品，没有A/B testing， 不知道是不是我答得太慢的缘故 心凉凉 -_-）
3. Q&A

详细说说analytical part，就是survey question那道题，地里有。facebook问卷库有很多questions, 每次随机发给用户，用户看到题目可以选择回答，跳过但不退出，或者不高兴答了直接退出。提供以下table，q_num是question的sequence number , 记录题目出现顺序。
Survey table
user_id|q_id|action {imp, ans, skip}|ts|q_num

（1）How do you evaluate which question is best answered?
就是求回答率最高的那道题吧，写完code, 我就主动提出concern，如果某些问题impression次数太少，对rate计算可能有影响，你想设置什么threshold for impression frequency吗？他说good question, 然后给我两个case, 数字是我刚编的，具体不记得了， 大概就是answer rate一样，但是分母很不一样， 你怎么判断这两种rate相同还是不同。。。我就说two sample proportion test，让写公式。

q1, ans1 = 1, imps1 = 10
q2, ans2 = 10, imps2 = 100

（2）If one user answered, say question #30, how would you decide which question comes next?
我就按conditional probability答的，说找相似users who also answered question #30, find the question with highest answer rate, 写了code，找他confirm。 他说你这里只考虑了其他所有非#30的questions，没有考虑sequence，我反应过来原来是要找所有答了#30的人下一题答题率最高的那道题。。。这里耗了好多时间。。。code写的不利索哎

（3）Given different algorithms of ordering survey questions, how do you evaluate which one is better? What metrics do you use?
我就说这个survey的objective应该是get more people engage and answer more questions. 可选的metrics可以有： average # of questions answers before exit, porportion of users answered all questions/completed the survey. 他又问以上两种方法， 你觉得哪种更能达到global optimal? 问到这里，心里很虚的说了第二种好。。。因为第二种updating with prior information而第一种只是randomizing没有prior information（欢迎讨论，难道是让要我做A/B testing了？？？我刚刚反应过来。。。当时想着这还是coding，不是product，就没想往experiment上说，后悔啊）。。。然后他又说，第二种只是在optimize下一道题的答题率，然而并不知道后面的答题情况，如果你想优化the length of questions answered (根据我前面说的metric)，就没法达到不是么。。。深深感觉自己给自己挖了个坑然后毫不犹豫得跳了下去。。。他说okay，forget about the previous two methods, can you think of a way to optimize your metric? 行吧，我说那咱就看length, 取答题最多的那些人的答题sequence, 准备写code来着，他说不用写code了只是了解一下思路。

然后就没有然后了，此时已经过去35分钟。。。直接进入QA环节。。。说好的product题呢？大概是因为这题答太慢了吧。

总体感觉题目不难，但是反应要快，要答得在点，这些还要多加练习。心累，明儿还要上班，要是想到别的再来补充吧。

最开始我不知道自己为什么挂了，因为所有的问题我都快速给出正解，并且附加了很多有创意的想法，甚至还教了面试官两招。
而当我被HR通知fail的时候，我震惊了，本来准备好的拒绝onsite的说辞都没用上。
可以说，我从来没有这么顺利而完美地interview过。
当然，之前bloomberg也有过一次，但是那次我反省出可能是我的态度太过鄙视，惹inerviewer不高兴了，所以我这次还特意小心翼翼的表现的很谦虚。
结果，败了
一直想不通为什么，直到我看到了这个视频，强烈推荐你们看一下：
https://www.youtube.com/watch?v=MfP-P8EHGBo
原因就是，当人家问你1+1 等于几的时候，只要回答2，千万不要附加其他的以证明自己懂很多。

我首先最大的误解就是FB的data scientist很高大上，以为他们有做或者至少懂得machine learning. 但是实际上并不是这样的。
所以当我在给对方讲解social network/ graph analysis 的时候，我就错了。 因为一个machine learning 算法，不管你有多强悍的英语表达能力，你不可能在三言两语内让对方明白此算法的真正用途，而对方只会以为你在吹牛。
哎，愚蠢了，当对方问我，“how you build the networks? what is edge and community?” 的时候，我就应该意识到，这伙计没ML背景。

好吧，回归正题，直接上答案
Two SQL coding questions, entry level的, 之前看的面经有人建议说一定要一边写一边解释，so I did. 但是感觉真的没必要，因为太简单了,还没说完呢，代码就写完了
Ads

| advertiser_id | ad_id | spend | Date | ... |
|---|---|---|---|---|

conversions

| ad_id | user_id | conversion$ | Date | |
|---|---|---|---|---|
| | | | | |

```
select advertiser_id , count(1) as total_spend_per_advertise
from Ads
group by advertiser_id //此处我面试的时候用的windowing 而非groupby，他就问我为什么，我说是在big data习惯，因为spark groupby有时候会比windowing慢
having date_diff(current_timestamp(),date) beween 1 and 30;

select a.count_of_advertisers_who_has_conversions/b.total_count_of_advertisers as
percentage_of_advertisers_who_got_conversions
from
(select count(distinct advertiser_id) as count_of_advertisers_who_has_conversions
```

from Ads
join conversions
on Ads.id=conversions.id) a, //此处问我逗号是啥意思，我说代表两个table的分隔，这里产生了cartesian join, 但是因为两个表格都只有一个数字，也无所谓，没必要单独生成table再计算。
(select count(distinct advertiser_id) as total_count_of_advertisers
from Ads
)b;

最后就是很case的问题， 问用什么metrics去衡量单个advertiser 的 广告效益。这里我貌似又讲多了，去讲什么生产profit, fixed cost, per unit cost, 没必要了，想想就是要么用conversion count rate, 要么用$conversion/$spend 就可以了。

第二部分ｃａｓｅ
问"What metrics would you use for the health of GROUPS"
因为之前代码节省了时间, 我这个问题就狂说，所以才讲到了network analysis...
"If we want add a new product to GROUPS, how do we evaluate its impact on the health”
这个就是ABtesting的东西了，我实话实说自己没有做过ABtesting，所以选择用student T和ANOVA解释如何一步一步建立样本，比较statistically difference。莫非说实话也减分，除了有医疗和网站背景的，没做过ABtesting很奇怪么。。。

**补充内容 (2018-4-21 04:50):**
还有，如果做个distribution of advertiser's total spend你觉得是什么样的。我说很难讲，可能是normal 或者是lognormal， 我偏向后者多一些，会有少数商家在facebooK的平台广告效益特别见效而愿意多spend。

**补充内容 (2018-4-22 05:56):**
感谢MK48指出，失败也极有可能是不懂AB testing，很有道理。不过此贴，LZ主要是想更多强调communication对DS的重要性，不管是在LZ目前的工作中，还是各种面试时，communication都是永远无法攻克的难题。。。

# User 9

第一次发帖，希望能惠及更多的同学 
TIMELINE:

- 3/6/2018 recruiter直接发邮件我问我要不要chat about ds opportunity at FB。当时觉得好假，不知道他哪里找来我邮箱的
- 3/9/2018 和recruiter 简短的打了一个电话，基本就是自我介绍。他介绍了FB这个组的信息，以及他找我是因为他们需要有quan背景的人。介绍了下一轮面试的流程
- 3/19/2018 我都以为他挂了电话就把我ignore了，楼主四月要抽签，就没有太放在心上了。结果这天突然把下一轮面试的详细信息发我了，当下楼主就认真起来了。本来约定四月第一周面，但是我说我那一周公司事情多，就改到了四月第二周（事实证明多一周准备实在是太重要了，楼主好多关键知识都是最后一周突然明白的  ）
- 4/12/2018 中午12：00面试，Interviwer不是约定的那个人。但是也没什么影响。一题SQL（两小问），接下来是产品。最后我问了三个小问题。12:55完美收官
- 当天下午六点多recruiter就电话我给feedback非常positive说比较outstanding. 邀请Onsite
- 20分钟后又电话问我现在visa. 我说抽签了在等结果， opt 2019夏天过期。recruiter 说抽中了可以transfer,但是FB 19年可以抽的名额满了。言下之意就是我现在抽不重，他们也不能确保给我抽。可能就不会安排onsite  劝我等到抽签结果，如果中了再找他们onsite

可想而知这一天过得像过山车一样，心情复杂。本来楼主背景不及地里大神，各种好的面试机会拿到手软，这次机会难得，楼主准备的也很辛苦，好不容易付出piad off, 不料卡在这个时间|身份点上。貌似以前在地里看到有人有类似情况。

喘口气～回来继续写真题。。。。。。。。

SQL：上来情景是一般fb会发给用户不同类型的message，有的是需要用户confirm的。后面巴拉巴拉一大串开头有点紧张也没有完全听明白。但是对于我们这种刷题型选手这不重要！只要上了table一下就能看懂了。不过大家还是要认真听intervewer说话,不要miss掉重要信息。例如：fb可能给同一个号码发送多个需要confirm的message. 这个就是个信息，你可能代码里需要写distinct！
我记忆里的信息，尽量还原，不保证一模一样。

table: SMS_INFO
column: { ds | country | carrier | user_number | message}
给大家解释下：
ds: datestamp (yyyy-mm-dd)
carrier: 手机的carrier
message: FB给用户发送的类型 （eg. message, friend_request, post, ....）

1） distribution of number of messages fb sent to each carrier in each country yesterday...
这个很显然了不多解释
提醒: 要会写yesterday的表达方式 （类似的还有几天前什么的....）
尽量think out loud, 一边写一边说。但是我知道SQL边说边写很清楚是很难的。
我的interviewer写完了他会让你go through你为什么那么写。所以注意练习下自己口语！表达一个问题的逻辑的能力

接下来第二小问。第二个table
table: Confirm_Info
column: {ds | number_point}
这里虽然叫另外一个名字，但其实就是第一个table里的phone number
这里可以看出，FB这个职位的面试套路满满，虽然想耍点心计，但是一般不会出现偏怪的东西。要细心！
2）confirm rate
具体不太记得了，当时立刻反应出来和friend request_acceptance_rate那种题目一样的，再次套路满满！
所以问了interviewer有没有threshold?比如24小时之内confirm的才算有效confirm?
面试官：good question! 你觉得threshold定多少？why
我说24 hour因为user cant check their phone from time to time, but after 24 hours, its easy for them to forget..
这里欢迎大家讨论其他的回答。
重点是没有标准答案，你自己要反应快，懂得弄一个合理的解释出来！
后来又问我几个细节的问题，为什么最后data type 是floating的啦，为什么用left join不用其他join啦

敲重点：
刷题， 刷题， 再刷题
每一个问题都会给你看一个result table 长什么样，所以你最后要保证column name, data type都一致
：细节细节再细节！
不会的，写错的，不要慌，面试上来十分钟大脑都会有点木讷，被紧张影响发挥最划不来了
面试官会稍微提示你哪里不太对，立刻反应就赶紧改就好了

FB人都满nice所以你自己也要放轻松～善于和面试官交流，不明白的赶紧问，一定要在写code之前clarify问题

沉着冷静，严肃活泼 : )

是confirm rate trend 之类的，trend就是group&order by time/datestamp啦大家都懂的套路～～～

**补充内容 (2018-4-14 03:56):**
CASE STUDY：
题目： 其实我也没有太听懂！

LOL意思是知道的就是FB以前在news feed上给大家推荐好友，现在想改成在side bar上了，你怎么evaluate 这个是不是work. that's it.

好了这是一个完美的套路的case study..

补充内容 (2018-4-14 03:57):
总体讨论路线： pick up metrics -> AB testing -> result after AB testing

补充内容 (2018-4-14 04:00):
1.选择metric: always keep our business goal in mind!!!这个feature到底是干嘛的: increase revenue/increase user engagement. etc...? 这道题，我觉得就是增加friend request & acceptance rate.
欢迎大家讨论

补充内容 (2018-4-14 04:03):

2.AB testing：上了udemy 和udacity上的AB testing课程。几大注意事项： sample group （no bias!）; sample size; run time...

补充内容 (2018-4-14 04:07):
3. Result: 及其喜欢问如果test蛮好的launch以后一个metric上升了但另一个下降了should we still launch?暂时大概和大家说下重点: what's the initial objective of this feature? why metric goes to this direction? Are they going to change along with the time if we launch ?

补充内容 (2018-4-14 04:08):
周五太累了改天继续说吧.....

补充内容 (2018-4-14 04:13):
说句题外话，虽然是个技术岗，每个人通过/挂的原因不同。努力和运气都有成分，但是我自己觉得SQL本来对大多数同学只要肯花功夫，肯定比写什么java python容易多。但是口语方面，因为这是个需要很多interaction 的岗位。我自认为口语强过大多数理科生，也有两年工作经验，但依然觉得想把一个问题有逻辑的说清楚，还需要锻炼。说的好绝对会加分，而且要想在美国职业真正有好的发展，这也是受益终身的技能。

补充内容 (2018-4-14 04:39):
发现了大神写的总结贴分享给大家 http://www.1point3acres.com/bbs/thread-326201-1-1.html

补充内容 (2018-4-16 22:32):
找大家一起购买a collection of Data Science take home challenge
http://www.1point3acres.com/bbs/ ... 7&page=1#pid3795236

补充内容 (2018-6-7 22:51):
后来昂赛的凉凉情况：http://www.1point3acres.com/bbs/ ... 9&page=1#pid4025759

## User10

4.29 面的facebook data scientist product. 有些紧张，觉得sql没答好。 大家加油，好好准备，放好心态，面试官人很好的。求大米和小伙伴一起准备ds product
SQL: 是market place 的题目：

Table 1: Commerce_user_actions
Date, sessionid (user click on MP tab inco or People search for commerce products), userid, event (surface_enter, click, surface_exit)

| Date | sessionid | userid | event |
|---|---|---|---|
| 2018-01-01 | session 1 | user 1 | surface_enter |
| 2018-01-01 | session 1 | user 1 | click |
| 2018-01-01 | session 2 | user 1 | surface_enter |
| 2018-01-01 | session 3 | user 2 | surface_enter |

Table 2: time_spent_per_session

Date
Sessionid
time_spent_sec

Q1. the avergae number of sessions/user by day for the last 30 days

```
select avg(ratio) from (
select count(distinct sessionid)/count(distinct userid) as ratio, day from commerce_user_action
where datediff(day, date, curdate())<=30
group by day)
```

Q2. Calculate the time spent distrbution on marketplace by users for a day? (x-axis is ts bucket, and y axis is number of uers

| ts-bucket (integer) | user_count |
|---|---|
| 0 | 200 |
| 1 | 100 |
| 2 | xxx |

ask about the distribution.

```
select intger(time_spent_sec) as ts_bucket,sum(cnts) as user_count from (
```

```
select date, sessionid, count(distinct userid) as cnts from commerce_user_action
group by date, sessioni) a. check 1point3acres for more.
join time_spent_per_session b
on a.date =b.date and a.sessionid = b.sessionid
and a.date =curdate()) c
group by ts_bucket
order by ts_bucket
```

Product:

there is a new feature of marketplace. when people log in facebook. There will be additional window introduce marketplace
- How would you test the feature (AB testing) and what metrics you are looking into?
- what are the potential bad impact of the new feature?
- Facebook cannot check the final results if the deal is success. They can tell the engagement through the number of messages per sesson.
What if # messenger per session goes down when we add the new feature?

one time or progressive
region, platform
denominator, numerator. more people come it but they don't actionally buy, the the number of sessions goes up but the number of messengers aren't

## User11

朋友去年内推的，12月初先跟HR phone screen，因为圣诞假期关系电面约到了1月。

上来先客套两句。对方是ABC, 毕业先在Capital One做了三年，然后跳到Facebook。然后让我简单介绍了一下自己。

接着就是SQL题，context是Facebook的marketplace，就是买卖东西的地方

scheme如下:

Table: commerce_user_actions
date [STRING] of format 'YYYY-MM-DD'
sessionid [BIGINT]
userid [BIGINT]
event [STRING]: surface_enter, surface_exit, click, first_scroll, message_send

第一问：For each day in the past 30 days, find the average number of sessions per user

```
SELECT date, COUNT(DISTINCT sessionid) / COUNT(DISTINCT userid) AS avg_num_sessions
FROM commerce_user_actions
WHERE date < DATE(NOW()) AND date >= DATE(NOW()) - INTERVAL '30 days'
GROUP BY date
```

答完第一题加入另一个table
Table: time_spent
date [STRING] of format 'YYYY-MM-DD'
sessionid [BIGINT]
time_spent [INT]

第二问：For each day in the past 30 days, find the number of users who had at least one session that was longer than 5secs

```
SELECT date, COUNT(DISTINCT userid)
FROM commerce_user_actions c
JOIN time_spent t
ON c.sessionid = t.sessonid
WHERE date < DATE(NOW()) AND date >= DATE(NOW()) - INTERVAL '30 days' AND
time_spent >= 5
GROUP BY date
```

第三问：Find the number of users who had at least one click everyday in the past 30 days

```
SELECT COUNT(userid) AS num_users
FROM
(SELECT userid, COUNT(DISTINCT date) AS num_days
FROM commerce_user_actions
```

WHERE num_days = 30 AND date < DATE(NOW()) AND date >= DATE(NOW()) - INTERVAL '30 days'
GROUP BY userid) temp

SQL题结束就是product sense题，依旧是Marketplace.
问：如果现在product team提议在用户每次进入Marketplace的时候都会跳出来一个prompt, 告诉用户在这个地区可以potentially reach多少人（比如说在湾区可以reach 500,000人），如何evaluate是否应该加这个prompt，a/b test用什么metrics.
追问，这个prompt可能有什么负面影响。

product sense题的答案我就不写了，开放问题大家可以自由讨论

## User12

数据表有两张，均为user log, 对于A表：每一次用户进入页面分派一个unique session id，用户离开则这一个session结束，期间用户的每一个行为都会生成一条记录；对于表B：记录一条session存在的时间。
A: date, session_id, user_id, act('enter', 'exit', 'post')--can be duplicated
B: date, session_id, time_spent --date and session_id are primary keys.

Q1：generate average number of session per user per day. 比如：
Date average_number_session_per_user
2018-09-09 30
Q2: generate number of user per time interval. in order to measure how many user is spending certain amount of time.
比如：
Time_spent number_of_user
0 4
1 6

产品：FB开发一个新产品，Pet page, 比如说Ins上有很多人专门给自己的宠物建一个账号，如果在FB上launch这个功能，让人们给自己的宠物建pet page，请问怎么measure这个产品
产品我就回答了[td]adoption perspective: adoption rate, growth rate such as WoW and the trend plot
DAU，MAU，etc

## User13

几周前面完facebook'的data scientist，感觉这个面试声誉不错的公司也没有特别的按照流程走吧，或者是我自己与Fb无缘。下面给大家具体说说我遇到的面试问题和流程，希望会有帮助。（一点背景：楼主现在亚马逊做data analyst II，已经工作3年多了，工作内容和fb ds的内容真的几乎完全一样，工具也基本都用过。。。）

1. 投完简历秒收到回复，recruiter太忙约到一周后，聊了30分钟双方基本信息。
2. 聊完recruiter又约了一周后的video screening并且发了一个提纲，大致是20mins Product, 20mins SQL/data process, 5min probability.
3. 面试之前一天下午收到另一个hr（应该是专门搞screening的）的电话，说我的面试官有事不能面，可不可以临时换一个人（印度人）。我想说无所谓吧就说ok。
4. 面试当天印度人如约而至，简单做了自我介绍（有口音），然后整场下来印度人问的题目大致是，35min product，10min sql。题目内容我记不清楚了，大体是这样的：

Product: Fb有一个security login notification service，大致就是说你假如上周在中国登陆fb，这周回美国登陆fb，系统会给你发封email问你这个人是不是你本人，不是的话可能有密码泄露赶紧改密码之类的。相信大家多少见过这种email。问题就是，Fb发布了这个功能之后，调查显示（survey result）对Fb的满意度下降了，这是为什么？我当时第一反应就是说太烦人了，我经常收到这种邮件所以应该是大家不喜欢fb总是发email。面试官说，有没有别的原因，我说大概有privacy的原因，就是大家可能认为fb自动获取了他们生活的location会有不安全感。面试官继续问，还能不能想出别的原因。。。我当时真的一直在瞎bb，想不出什么别的了。然后灵光一现，想到应该是survey data的问题，就说会不会是survey result的问题。印度人点头，然后我说，两组survey的人不同，可能一组人本身就很喜爱Fb所以打分高，另一组虽然有了新功能但是本身有些sample人们不喜欢Fb，所以给的结果就有bias。面试官继续问，那有没有别的问题。我问他，这是什么样的survey。面试官说，monthly through email。然后我问，survey问题month over month有改变吗？面试官说，没有改变。我说，大概是什么样子的问题？面试官想了半天说，for example，how do you like facebook? rate from 1 to 5. 我当时就说。。。那这种survey完全不能代表这个security login notification的表现啊，这个survey太广了。面试官说，你觉得还有什么具体问题？我就说，一个月里fb可能不止launch这么一个login notification feature，这个survey结果可能是别的feature造成的。面试官说，假如我们这个月全公司就launch了这一个feature呢。我说，那可能是sample人群不同，就是比如两组samle里有的是我们这种年轻人多就不太会care，用没用这个feature都给高分，另一组sample里这种年轻人就偏少导致分数不管怎样都会低一些。面试官接续说，还有没有别的问题。。。我又绞尽脑汁想了个说因为是month over month所以还会有seasonality的问题，可能比如12月底圣诞节大家普遍开心就给分高一点，1月底上班了什么的大家有压力就会比较picky一点给分低一些。。。（以上我写的很流畅但是其实当时我是整个一直在不停的瞎bbbb，我英文也一般，还省去了一些没用的小对话之类的，面试官当时已经是很不耐烦的状态了，当时我记得已经在25-30分钟了吧，总感觉没有抓住面试官真正想要问的点，当时心情真的是很慌张了）然后面试官问，那你觉得应该怎样做这个survey。我说，直接发一个survey target on this feature. 面试官说，不能改survey内容也不能增加survey题目。我说，那我们可以根据用户engagement数据来找到那些一个月里除了这个security login之外没用过任何其他新feature的人，只用他们做sample就可以排除其他feature的干扰。面试官说，这不是一个数据题，我们只聊产品，假设你不能看数据，就从产品的角度上来回答。。。我此时基本是懵逼状态了，不给数据那我真的想不出来什么了。然后我就不太记得了，好像中间还有些别的很多，为什么，你觉得应该怎么做，之类的问题，我都是绞尽脑汁答的，然后感觉答的都乱七八糟的。一直到最后面试官就是陪着笑那种，说，我觉得你说的不错（很场面话），我们来做个sql题。这时候大概已经在30-35分钟了。

SQL：sql题目比较简单：一个广告table，每行primary key是时间和广告ID，列是timestamp

，adsID（广告ID），publisherID（广告商名字），还有广告价格。另一个table是该广告有多少人看见，多少人点击。列是timestamp，adsID，#views，#clicks 吧，具体的我想不起来了。。。反正就是很基本的求某天某广告商的convertion rate。但是我当时一直脑子里在想上面那个题，我感觉肯定有个东西没说上来，这个不补救的话肯定要挂了。所以sql就随便做了下。就秒写了两个小sql，一个是分子的，一个是分母的。然后step by step给面试官简单的说了下相除是convertion。他说好的，你能不能用同一个sql写，我说ok。然后写了个join，此时脑子里还在想挽救上面的那个产品题。然后这里出现了事后我认为的误会：（以下sql是编的，为了突出问题编了一下。具体我已经不记得了）

我写的：select a.xx, b.yy from table_ads a left join table_clicks b on a.extract(date from timestamp) = b.extract(date from timestamp) where a.extract(date from timestamp) = current_date ..... 然后我又把from改了一下，改成了这个：from (select * from table_ads where extract(date from timestamp) = current_date) a left join table_clicks b on a.extract(date from timestamp) = b.extract(date from timestamp). 面试官跟我聊了几句，问我为什么要改，这两种写法的差别在哪里。我这个人解释东西比较嘴笨，还是用英文。。你们体会一下这个区别怎么给别人讲。。。我讲了半天还在屏幕上拿刚才写好的sql举例子啥的，面试官表示，I don't understand... 我就当时真的很慌了，语言越发不利索。最后面试官说，其实这两个是没有区别的。。我说，有区别啊。。。。面试官说，这道题里你用哪种都是一样的，我说，对。面试官很无语说，你刚不是还说有区别吗，还给我解释了半天，虽然我没听懂。我说，sql有区别的但是对这个题的话最后的结果是没区别。。。说完我自己都觉得自己很像见风使舵。。。然后最后的时候我猛然发现他求的是某广告商（publisherID）的convertion，而我前面写的是group by ads_id（广告的convertion），然后我说我发现了一个小错误能不能改一下。其实当时屏幕上的sql已经有点乱七八糟了因为我前面举例子啥的改来改去。。最后面试官说超时了，没关系他已经了解了。当时大概已经过了35分钟了。

Q&A：面试官说，你还有什么问题。我说，我觉得自己发挥的不好，你知道再申ds的话要freeze多久吗？他说，我觉得你发挥的很好啊（就是这里给了我后面无尽的等待）。然后我又问了一个关于选组的问题就结束了。

5. 面试结束之后我等了一个周，给schedule video的hr发了邮件问结果怎么样，她过了一天回复说，要我问原先找我的recruiter。我就给原来的recuiter发了一封email问。recuiter回复说一直没有联系到我的面试官所以没有结果，他会抓紧问。于是我又等了一个周，还是没动静。其实本来到这里我就默认自己挂了，但因为最后面试官跟我说他觉得我发挥得很好，我就内心觉得可能还有戏。。于是又发了email给recruiter问，他说，还没得到面试官的回复，他明天就去找面试官问。然后我又等了2天没消息。。就找了个一个fb的朋友帮我问问。。大概是fb的朋友为自己公司这个缓慢的速度赶到羞愧，就直接去问了recruiter的老板。。然后他老板说，不能告诉我朋友结果但是这个recuiter已经告诉我结果了。。。可是并没有。。我跟我朋友分析大概是recuiter撒谎了，告诉他老板他已经联系过我了，其实没有，然而也不知道为啥。然后又过了一天，我突然收到recuiter的email说他之前一直发给了一个错误的邮箱地址。。。我也是醉了我催他那些email他直接回复不就好了吗。。。总之他在第二个周终于找到了面试官问面试官说didn't pass.

总结了一下整场下来的心得：

1. screening这个面试不像on site有多人，所以全靠面试官。假如临时给你换人，尽量不要接受。因为我自己觉得我这个面试官可能没有准备什么很充分的问题，临时被拉来，所以一直在这个login product的问题上扩展。

2. 要多练习开脑洞的产品题，虽然我在亚马逊产品组呆了1年多（之前换组）但还是觉得对中国人来说可能产品问题开脑洞系列是最难的。

3.勤联系recuiter，不要默认自己没过。像我这个面试官就是面完了2个礼拜都没给结果，还是我催

着recuiter去催他的。
3. 勤联系recuiter，万一我当时真的通过了，他给我发到错误的email我也没有一直回复他的话，搞不好就错过了这个机会。
4. 产品题放在最前面答的不好可能会给后面带来压力，不要妄想挽救肯定没戏的，要focus在后面的题，听清楚题意，不要产生误会。
（5. 不管怎样还是感觉印度人有点坑。。）

祝大家好运！

1) get the highest answer rate question
2) how to dynamically change the order of the questions showing to the users: 也就是用户如果回答或者跳过了一个问题，那么下一个问题应该如何分配给user，来优化用户回答问题的概率

change of the order of friends your might known and show it on the newsfeed instead of the side bar on the right, how to evaluate its performace
follow up question: if friends question goes up, but engagement goes down, should we launch this product?

之前求地里的大佬内推， 收到了HR的邮件， 回答了十几个问题，然后跟HR约了时间，说是15分钟的电话聊天。
刚刚通完电话了， 跟地里之前看到的一个兄弟说的东西一模一样！贴一下流程：
没有introduce experience
直接问
1. Why data scientist and facebook
2. Tell me an experience when data influence decision
然后可能我回答的比较长， 他没有问我'what do you think a data scientist will do at FB'， 而是直接跟我说了一通，然后说跟deep learning/machine learning完全不沾边（跟之前地里一位说是FB这个职位的面试官说的东西一模一样），比较注重product thinking。
然后是几个high level 的sql题目：（跟之前那个面经一样）

- Order by 默认排序顺序是什么： ASCENDING.
- 统计一个序列中不重复的数字的个数： Count(distinct ...)
- 左表是大表，右边是小表，要保留左边所有的条目应该用什么join的方式： left join
- 10 + null + 5 + 3 等于多少: null

挺水的毕竟是HR问但是好像我有同学在HR这轮挂了。。。所以，感觉你还是得把自己说的比较偏data analysis/BA 这个方向，不要太强调machine learning啥的，然后sql题目不要错吧！

## User 13

I interviewed for a DS position at FB. My phone screen consisted of one fair straightforward SQL question (I think it required a join and an aggregation), and then one longer analytics scenario. You could think through the products that FB has (FB, Insta, etc) and think about what sorts of decisions have been made at various points (frequency of ads to show, which friends to recommend, which ads to show, etc) and think through what kinds of data would you collect and how you would Analyze it to

make the data. What would you collect the data? What metrics exactly would you use? Why are those the best? They might say ok, you do that experiment and your results are ___. What would you conclude? etc.

Questions:

SQL and Business case study:

The interview consists of coding (SQL or Python) as well as a few technical mathematical questions (mainly probability, using various aspects with a deck of cards) and some questions about how mathematics could be applied to the real world (how could graph theory be Used to verify users' information?).

Advertiser and user spend, basic SQL questions to pull number advertisers, ROI for each , etc?

How can Facebook figure out when users falsify their admitted schools?

SQL/R/Python stuff and define the right metric for certain FB products

SQL queries with basic group by, self joins and inner queries. The problem could be solved by analytical queries Self joins And inner queries. The problem could be solved by analytical queries

Business problem: Implementing creation of new reaction like happy, sad etc

Coding:

Given an list A of objects and another list B which is identical to A except that one element is removed, find that removed element.  An list

Leetcode word search 1 and word search 2

I got offered:
122k base+
10% bonus
150k Rsu /4 years

I accepted offer.

第一轮就挂了 就算给后面人一个提醒吧
一共三题
第一题是关于machine learning cluster方面的东西 我答得一般
第二题 是mysql 我之前一直在coding mysql 不熟，所以估计是挂的原因
第三题 是对facebook 产品的意见和建议 也感觉自己说的不号
总结一下就是facebook 还是要多了解他的产品 最后祝大家好运

# 1. Workplace

Q1: 提供的table: 一个是 region (like Asia, Europe) | country | company_id, 一个是 company_id | number_of_onboarded_employee. calculate the average number of onboarded employees per country and per region.

```
SELECT B.region, B.region_avg,
A.country, A.country_avg
FROM
(
SELECT t1.country, AVG(t1.number_of_onboarded_employee) as country_avg,
COLLECT_SET(t1.region) as region
FROM t1
JOIN t2 on t1.company_id=t2.company_id
GROUP BY 1
) A

JOIN

(
SELECT t1.region, AVG(t1.number_of_onboarded_employee) as region_avg
FROM t1
JOIN t2 on t1.company_id=t2.company_id
GROUP BY 1
) B ON A.region=B.region
```

Q2: 表1: company id, ds, company_create_time, country, region
ds是每天的日期，每天产生一个表，然后append到最后
表2: company_id, date_joined, user_id
userid是员工，datejoined是公司让员工用这个产品的时间

1) 某个地区每个国家在最近30天内有多少公司注册了

```
SELECT country,
COUNT(DISTINCT company_id) AS cnt
FROM table1
WHERE region='Asia'
AND company_create_time BETWEEN curdate() AND curdate()-30
GROUP BY 1
```

2) how many users do companies on board by the end of their first week (make this a daily tracking metric)

```
SELECT tb.day,
AVG(cnt) as average
FROM
(
SELECT A.company_id,
DATEDIFF(B.date_joined, B.date_joined+7) as day,
COUNT(B.user_id) as cnt
```

```
FROM table1 AS A
JOIN table AS B
ON A.company_id=B.company_id
WHERE ds BETWEEN B.date_joined AND B.date_joined+7
GROUP BY 1,2
) tb
GROUP BY 1
```

# 2. Instagram video chat

| date | caller_id | receiver_id | duration | caller_country | receiver_country |
|---|---|---|---|---|---|
| '2018-07-01' | 1234 | 7567 | 63.4 | 'ES' | 'ES' |
| '2018-07-01' | 1234 | 3669 | 50.8 | 'ES' | 'ES' |
| '2018-07-01' | 1234 | 8998 | 0.0 | 'ES' | 'PT' |

1) for people who used the feature *for the first time* on 2018-05-05, how many people used the feature X days later? How do you tell the 7th day retention is good or bad? (cohort analysis when x=7)

```
SELECT datediff(DATE, '2018-05-05', t1.date) as Xth_day,
COUNT(DISTINCT t1.uid) as cnt
FROM
(
SELECT date, caller_id as uid
FROM table
UNION ALL
SELECT date, receiver_id as uid
FROM table
) as t1

JOIN

(
SELECT t1.uid, MIN(t1.date)
FROM
(
SELECT date, caller_id as uid
FROM table
UNION ALL
SELECT date, receiver_id as uid
FROM table
) t1
GROUP BY 1
HAVING MIN(date)='2018-05-05'
) t2 ON t1.uid=t2.uid
GROUP BY 1
ORDER BY 1
```

2) This new feature is only available to specific users for testing, and for each day the features exposed to the users as in the following table: users who have access to this feature might not use it at all. (* receiver_id and caller_id in the upper table are all user_id). If a user uses this function for at least 3 seconds for

one time a day, then we call this users as an active user of this day. What's the percentage of active users among all the users who have access to the feature for each day?

```
SELECT date,
SUM(active) / COUNT(*) as rate
FROM
(
SELECT date, user_id,
IF(duration>=3, 1, 0) AS active
FROM table
) temp
GROUP BY 1
```

# 3. Marketplace

Table 1: Commerce_user_actions, having columns as Date, sessionid (user click on MP tab inco or People search for commerce products), userid, event (surface_enter, click, surface_exit)

| Date | sessionid | userid | event |
|---|---|---|---|
| 2018-01-01 | session 1 | user 1 | surface_enter |
| 2018-01-01 | session 1 | user 1 | click |
| 2018-01-01 | session 2 | user 1 | surface_enter |
| 2018-01-01 | session 3 | user 2 | surface_enter |

Table 2: time_spent_sec with sessionid and timespent

(Sessions: Date | Session_id | User_id | Action (enter/click/send/exit)
Time: Date | Session_id | Time_spent (s) )

Q1: Calculate the average number of sessions/user per day for the last 30 days

```
SELECT COUNT(session_id) / COUNT(DISTINCT user_id)) as average
FROM table 1
WHERE DATEDIFF(curdate(), date)<=30
```

Q2: Time distribution of each user. What may the distribution look like?

```
SELECT time_spent,
COUNT(user_id) as cnt
FROM table 1 as A
JOIN table 2 as B
ON A.session_id=B.session_id
GROUP BY 1
ORDER BY 1
```

Q3. # of users who at least spent more than 10s on each session

```
SELECT COUNT(DISTINCT user_id)
FROM session
```

```
JOIN time
ON session.sessionid = time.sessionid
GROUP BY time.session_id
HAVING MIN(time_spent) > 10
```

Q4: Average time spent on session 1 per user within the last 30 days

```
SELECT IFNULL(AVG(time_spent), 0)
FROM session
LEFT JOIN time
ON session.sessinid = time.sessionid
WHERE session.sessionid = '1'
AND DATEDIFF(curdate(), date) <=30
GROUP BY user_id
```

Q5: Plot the histogram of avg(time_spent). How do you know within certain time period, how many people are in there?

```
SELECT tb.avg_time,
COUNT(DISTINCT tb.user_id)
FROM
(
SELECT userid, AVG(time_spent) as avg_time
FROM session
JOIN time ON session.sessinid = time.sessionid
WHERE session.sessionid = '1'
AND DATEDIFF(curdate(), date) <=30
GROUP BY user_id
) tb
GROUP BY 1
ORDER BY 1
```

Q6: daily active user for the past 30 days (event with open session/end session/scroll down/first click/send message)

i. first define DAU

- session > 5s
- scroll down或者first click才算，因为只打开一个session然后time out或者就quit的话不应该算，然后send message表示至少有一个click，所以send message的session肯定都有first click，所以最后还是选择scroll down和first click的

ii. sql

```
SELECT A.date,
COUNT(DISTINCT A.user_id) as DAU
FROM table1 AS A
JOIN table 2 AS B
ON A.session_id=B.session_id
WHERE B.session_time>=5
AND A.event IN ("scroll down", "first click")
GROUP BY 1
```

Product Sense
marketplace tab里加一个promoter “sell your item”

1) Estimate the distribution of time spent on market place
2) 会有什么影响 (What is the benefit of launching Call to action (e.g., learn more, buy)?)
3) 怎么来估计这个影响(what metrics)
   Goal is to encourage people to sell products
   Metrics:
   # of users who sell their products
   # Conversion rate (# people clicked yes/# people viewed)
   # Monthly/daily active users
   # Engagement/Time spent on browsing marketplace
4) The message_sent/session drops 10% month by month. What is the reason?

- Data accuracy/spam/outlier
- area effects, some regions drops, while some regions increases. Need further investigation on this problem.
- Replied Seasonality effect (look at historical data), time span(two months data vs. six months data), maybe buyer will not use facebook messenger to leave a note to buyers, competitors, does they show similar trend?

5) marketplace 想给user 推送 item recommendation 在newfeed 里面，要怎么设计 algorithm 去推荐? 用什么data 去train algorithm（有click 和purchase 的data 用哪一个好）? 有了algorithm 怎么去测试? launch 之后怎么update 既有的algorithm

- Estimate the probability of purchasing an item based on
  - Frequency of purchasing items under the same category
  - Price of past purchases
  - Whether from user’s friends or group
  - Whether newly added products (clothes, accessories, electronics, trendy products, books)
- Use the puchase data to train
- Test the algorithm: precision, recall, accuracy, F-score
- Launch: AB Test
  - metrics: ratio of actual purchase (time of purchases / time of views)
  - random select control and treatment group

6) How would you define meaningful Marketplace DAP/DAU?
   - DAU: time of views / session time spent>5s / scroll-down / first click
   - DAP: average amount spent by active user
7) We are creating an upsell CTA (call to action) - when people click into marketplace tab, we also prompt them to “sell something”. Why should we as in Facebook do this? What metrics would you look at in order to decide if we should launch or not?
   - increase engagement and interactions among users
   - metrics: CTR or conversion rate (# of actual sell / # of views)

# 4. confirmation text

FB会发给用户confirmation text。给了一个表里面有date | country | carrier | uid | status（confirmed or not）。问

1) 找出每个国家每个carrier的confirmation 总数。

SELECT country, carrier,

```
SUM(IF(status='comfirmed',1,0) as total
FROM table
GROUP BY 1,2
```

2) 过去30天的confirmation rate。

```
SELECT
SUM(IF(status='comfirmed',1,0) / COUNT(*) as rate
FROM table
WHERE date BETWEEN '2019-12-01' AND '2019-12-31'
```

6. sms_message (fb to users)

```
date     |country|cell_number |carrier  |type
2018-12-06 |US     |xxxxxxxxxx|verizon | confirmation (ask user to confirm)
2018-12-05 |UK    |xxxxxxxxxx|t-mobile| notification
confirmation (users confirmed their phone number)
|date  |   cell_number |
(User can only confirm during the same day FB sent the confirmation message)
```

1) Yesterday how many confirmation texts by country.

```
SELECT country,
SUM(IF(type='confirmation', 1, 0)) as cnt
FROM sms_message
WHERE date = curdate() - 1
GROUP BY 1
```

2) How many requests fb sent to each carrier yesterday?

```
SELECT carrier,
SUM(IF(type='confirmation', 1, 0)) as cnt
FROM sms_message
WHERE date = curdate() - 1
GROUP BY 1
```

3) Number of users who received notification every single day during the last 7 days.

```
SELECT tb2.cell_number
FROM
(
SELECT tb1.date, tb1.cell_number,
IF(notification_cnt>0, 1, 0) notification
FROM
(
SELECT date, cell_number,
SUM(IF(type='notification', 1, 0)) as notification_cnt
FROM sms_message
WHERE date BETWEEN '2020-01-01' AND '2020-01-07'
GROUP BY 1,2
) tb1
) tb2
HAVING SUM(tb2.notification)=7
```

4) On dec 06th, overall confirmation rate.

```
SELECT
SUM(IF(B.cell_number is NULL,0,1)) / COUNT(A.*) as rate
FROM sms_message AS A
LEFT JOIN confirmation AS B
ON A.date=B.date
WHERE A.date='2019-12-06' AND B.date='2019-12-06'
```

FOLLOW-UP:

1) If the confirmation rate decreased by x%, what might be the reason?
   - significant & size (statistical & practical)
   - data accuracy
   - outlier/spam (bot asking for verification a lot of time)
   - known & unknown technical problems (internal from the tech department or external reports/complaints)
   - slice it into category (country, carrier, time of the day)
   - by denominator and numerator (# of sent, # of confirmed)

2) Assume the number of messages FB sent don't change, but confirmation rate decreased by x%, why?
   - spam and bot users
   - carriers tech problem
   - email/carrier update FB messages into spam and user do not read it

3) Assume carrier is the reason for confirmation decrease, how to find which carrier?

```
SELECT carrier,
SUM(IF(B.cell_number is NULL,0,1)) / COUNT(A.*) as rate
FROM sms_message AS A
LEFT JOIN confirmation AS B
ON A.date=B.date
GROUP BY 1
```

# 5. Compose

Q1 - Event: ds | host_id | action | event_id | interface
action 有 start, add_location, upload_photo, publish
interface 有android, iphone

1) how many events were published yesterday, on each day interface?

```
SELECT interface,
COUNT(event_id) as cnt
FROM Event
WHERE ds=curdate()-1
GROUP BY 1
```

2) What percent of events that people start creating get published?

```
SELECT
```

```
SUM(IF(action=”publish”, 1, 0) / SUM(action=”start”, 1, 0) AS percent
FROM Event
```

Q2 - 给了一个table composer，3 columns: userid | event | date, event包括enter/post/cancel (enter就是开始在composer里面写内容，cancel就是开始编辑但是没有post而是终止了)

1) what is the post success rate for each day in the last week?

```
SELECT date,
SUM(IF(event = 'post’,1,0)) / SUM(IF(event = 'enter',1,0)) as rate
FROM composer
WHERE datediff(day, date, current_date) <= 7
GROUP BY date
ORDER BY date
```

2) 每个人每天的success rate是多少

```
SELECT date, user_id
SUM(IF(event = 'post’,1,0)) / SUM(IF(event = 'enter',1,0)) as rate
FROM composer
GROUP BY 1,2
```

3) 在第一题的基础上，又给了一个table：user，4 columns: userid | date | country | dau_flag{0, 1}。其中dau_flag表示daily active or not. What is the average number of post per daily active user by country today?

```
select country,
ifnull(num_post/num_user, 0) as avg_post_today
from
(
select country,
count(distinct user) as num_user,
count(userid) as num_post
from user AS A
join composer AS B
on A.userid = B.userid AND A.date=B.date
where A.dau_flag = 1
and A.date = curdate()
)
group by country
```

**Product Sense**

1）上面题中metric - average number of post per daily active user 突然从3下降到2.5，有哪些可能的原因，并且解释每个原因

- data accuracy? spam or outliers?
- post decrease or DAU increase or both?
- DAU: seasonality (external), experiments or new features (internal)
- post: change in UI/feature, time-pattern
- slice users into categories

2）add the composer tab from facebook to insta, what metric to evaluate

- post success rate, click through rate
- engagement: time people spent, session length/interval, DAU/MAU

- long-term: retention rate, revenue

# 6. SPAM

Table: user_actions
ds (STRING) | user_id (BIGINT) |post_id (BIGINT) |action (STRING) | extra (STRING)
'2018-07-01'| 1209283021 | 329482048384792 | 'view' |
'2018-07-01'| 1209283021 | 329482048384792 | 'like' |
'2018-07-01'| 1938409273 | 349573908750923 | 'reaction' | 'LOVE'
'2018-07-01'| 1209283021 | 329482048384792 | 'comment' | 'Such nice Raybans'
'2018-07-01'| 1238472931 | 329482048384792 | 'report' | 'SPAM'
'2018-07-01'| 1298349287 | 328472938472087 | 'report' | 'NUDITY'
'2018-07-01'| 1238712388 | 329482048384792 | 'reshare' | 'I wanted to share with you all'
Table: reviewer_removals （真SPAM）
ds (STRING) | reviewer_id (BIGINT) | post_id (BIGINT) |
'2018-07-01'| 3894729384729078 | 329482048384792 |
'2018-07-01'| 8477594743909585 | 388573002873499 |

1) how many posts were reported yesterday for each report Reason?

```
select extra, count(distinct post_id)
from user_actions
where ds = curdate() - 1 and action = "report"
group by extra
```

2) What percent of daily content that users view on Facebook is actually Spam?

```
SELECT u.date,
COUNT(DISTINCT r.post_id)/COUNT(DISTINCT u.post_id) as spam_percentage
FROM user_actions u
LEFT JOIN reviewer_removals r
ON u.post_id = r.post_id
WHERE u.action = 'view'
GROUP BY u.date;
```

3) How to find the user who abuses this spam system?

```
SELECT A.user_id,
SUM(IF(A.action=”report”, 1, 0)) as report_cnt,
SUM(IF(B.post_id IS NULL, 0, 1)) as spam_cnt
FROM user_actions AS A
LEFT JOIN reviewer_removals AS B
ON A.post_id=B.post_id
```

# 7. Friend

Q1: Table【Friending】
time = timestamp of the action
date = human-readable timestamp, i.e, 2018-01-01

action = {'send', 'accept'}
actor_id = uid of the person pressing the button to take the action
target_id = uid of another person who is involved in the action

- Define how long you have to wait before a friend request is considered rejected (e.g. 1 week) → find the average number
- Here a user may send multiple request to a same user at different time

1) 某日，有多少人发好友申请，有多少人接受好友申请

```
SELECT
SUM(IF(action='send', 1, 0)) as send,
SUM(IF(action='accpet', 1, 0)) as accept
FROM friending
WHERE ds='2020-01-04'
```

2) 每天有多少人成功交友（双向的），要group by date

```
SELECT A.ds,
COUNT(DISTINCT A.action_id) as cnt
FROM
(
SELECT ds, action_id, target_id
FROM friending
WHERE action='send'
) A

JOIN

(
SELECT ds, action_id, target_id
FROM friending
WHERE action='accept'
) B
ON A.target_id=B.action_id AND A.ds=B.ds
GROUP BY 1
```

3) What was the friend request acceptance rate for requests sent out on 2018-01-01?

- 如果multiple request不需要改动

```
SELECT
SUM(IF(action="accept",1, 0)) / SUM (IF(action="send",1, 0)) as rate
FROM Friending
WHERE d_date="2018-01-01"
```

- 如果multiple request需要被当成一次

```
SELECT SUM(IF(action="accept",1, 0)) / SUM (IF(action="send",1, 0)) as rate
FROM
(
SELECT actor_id, target_id, action
```

```
ROW_NUMBER() OVER(PARTITION BY actor_id, target_id, action ORDER BY actor_id) as
index
FROM Friending
WHERE d_date="2018-01-01"
) tb
WHERE tb.index=1
```

4) Find friend acceptance rate trending

```
SELECT d_date,
SUM(IF(action="accept",1, 0)) / SUM (IF(action="send",1, 0)) as rate
FROM Friending
GROUP BY 1
ORDER BY 1
```

5) 如果action中有unfriend，要求计算每个人的好友. 如何判断两个人是不是好朋友

```
SELECT A.action_id,
COUNT(DISTINCT B.action_id) as friend_cnt
FROM
(
SELECT action_id, target_id
FROM friending
WHERE action=”sent”
AND (action_id, target_id) NOT IN
(
SELECT action_id, target_id
FROM friending
WHERE action=’unfriend’
)
) A

JOIN

(
SELECT action_id, target_id
FROM friending
WHERE action=”accept”
AND (action_id, target_id) NOT IN
(
SELECT action_id, target_id
FROM friending
WHERE action=’unfriend’
)
) B ON A.action_id=B.target_id AND A.target_id=B.action_id
```

Q2: Recommend pages your friends liked.


TABLE1: Friends {user_id, friend_id}
TABLE2: Page {user_id, page_id}

Write an SQL query to make recommendations using pages that your friends liked. The query result should not recommend the pages that have already been liked by a user.

```
SELECT A.user_id, B.page
FROM Friends AS A
JOIN Page AS B
ON A.friend_id=B.user_id
WHERE (A.user_id, B.page) NOT IN
(
SELECT A.user_id, B.page
FROM Friends AS A
JOIN Page AS B
ON A.user_id=B.user_id
)
```

Q3: Friend Requests I: Overall Acceptance Rate
Write a query to find the overall acceptance rate of requests rounded to 2 decimals, which is the number of acceptance divided by the number of requests.

Table: friend_request

| sender_id | send_to_id | request_date |
|---|---|---|
| 1 | 2 | 2016_06-01 |
| 1 | 3 | 2016_06-01 |
| 1 | 4 | 2016_06-01 |
| 2 | 3 | 2016_06-02 |
| 3 | 4 | 2016-06-09 |

Table: request_accepted

| requester_id | accepter_id | accept_date |
|---|---|---|
| 1 | 2 | 2016_06-03 |
| 1 | 3 | 2016-06-08 |
| 2 | 3 | 2016-06-08 |
| 3 | 4 | 2016-06-09 |
| 3 | 4 | 2016-06-10 |

For the sample data above, your query should return the following result.
Note:

- The accepted requests are not necessarily from the table friend_request. In this case, you just need to simply count the total accepted requests (no matter whether they are in the original requests), and divide it by the number of requests to get the acceptance rate.
- It is possible that a sender sends multiple requests to the same receiver, and a request could be accepted more than once. In this case, the ‘duplicated’ requests or acceptances are only counted once.
- If there are no requests at all, you should return 0.00 as the accept_rate.

```
SELECT
ROUND(COUNT(DISTINCT B.accepter_id) / COUNT(DISTINCT A.sender_id), 2) AS rate
FROM friend_request AS A
LEFT JOIN request_accepted AS B
ON A.sender_id=B.requester_id AND send_to_id=B.accepter_id
```

FOLLOW-UP

1) How would you conduct an experiment to test if a change in facebook app is effective and what metrics will you look at?
   AB Test (before launch) or Observational Research (after launch)
   Metric: engagement
2) What metric would you show small businesses if you were trying to have them sign up for Facebook Ads
3) If you implement a new feature to FB how would you measure the success

- 先搞清楚产品的目标是什么，你可以问面试官：what is the goal of this product? 从目标出发，先定义衡量的metrics，然后就是做实验了，根据实验结果来判断是否成功。
- 在定义metrics的时候要想的全，哪些metrics对你的目标是最重要的，另外也不要忘了定义counter metrics，定义metrics想3个最重要的就可以了。我下面用一个例子来说明这类问题的思路。还是以社交网络为例子，他们改变friends recommendation 算法，希望用户可以通过这个功能多加好友，如何衡量这个feature是不是成功呢。我们先明确这个功能的目标，是希望多加好友，加了好友以后，希望在平台上有更多的交涉。那么这个时候的metrics可以应用funnel的思路来定义：好友增加率（friends request/accept rate），月活量（monthly active user），参与度(engagement），这些方面来定义metrics，尽量想到至少3个。然后做ab test，比较两个group的差别。
- 做完实验后面试官会有一些follow up问题，比如metric a 增加，metric b减少. 比如我们发现monthly active user(mau)增加了5%， 但是engagement 减少了3%，那么这个产品是好是坏呢。这个时候engagement就是我们的counter metrics，因为我们不仅希望加了好友以后，每个人的connection增加了，我们更看重是不是有意义的connection， 所以engagement也是一个很重要的指标。一般这种题目可以从短期vs长期效应来考虑。比如我们的mau增加了，虽然短期的engagement减少了，但是长期看的话，由于每个人的connection增加了，由于network effect，大家相互影响，engagement长期看也许会增加的，这个时候就要去衡量long term effect了。
- metric a增加了2%，那么这些情况下是否应该launch这个产品呢。假设我们的metrics都正向增加了，比如2%，那么如何判定这个2% 是个好的增长，可以launch to everyone? 这个时候可以往量化方面想，比如2%的增加对应了多少popualtion，这个2%的增加带来的潜在revenue是多少，如果是好几百万的话，那即便2%，也是个很好的提升！

Q4: Friend Requests II: Who Has the Most Friends

| requester_id | accepter_id | accept_date |
|---|---|---|
| 1 | 2 | 2016_06-03 |
| 1 | 3 | 2016-06-08 |
| 2 | 3 | 2016-06-08 |
| 3 | 4 | 2016-06-09 |

Write a query to find the person who has most friends and the most friend number. For the sample data above, the result is:

| id | num |
|---|---|
| 3 | 3 |

Note:

- It is guaranteed there is only 1 people having the most friends.
- The friend request could only been accepted once, which mean there is no multiple records with the same requester_id and accepter_id value.

```
SELECT requester_id,
COUNT(accepter_id) as cnt
FROM table
GROUP BY 1
```

```
ORDER BY 2 DESC
LIMIT 1,1
```

Follow-up: In the real world, multiple people could have the same most number of friends, can you find all these people in this case?

```
SELECT requester_id,
COUNT(accepter_id) as cnt
FROM table
WHERE (requester_id, COUNT(accepter_id)) IN
(
SELECT requester_id,
MAX(COUNT(accepter_id)) as max
FROM table
GROUP BY 1
)
GROUP BY 1
```

```
SELECT tb2.requester_id, tb2.cnt
FROM
(
SELECT tb1.requester_id, tb1.cnt,
RANK() OVER(PARTITION BY tb1.requester_id ORDER BY tb1.cnt DESC) as rank
FROM
(
SELECT requester_id,
COUNT(accepter_id) as cnt
FROM table
GROUP BY 1
) tb1
) tb2
WHERE tb2.rank=1
```

Q5: Share with friends

1) 从 User1| user2 如何找到 top 10 users with most friends

```
SELECT user1,
COUNT(DISTINCT user2) as friend_cnt
FROM table
GROUP BY 1
ORDER BY 2 DESC
LIMIT 10
```

2) 第二个问题是加了个Table2 sender_id | recipient id | content_type| date 问如何找到2019年每对friend之间 share的次数

```
SELECT tb.user1, tb.user2,
SUM(share_cnt) as share_cnt
FROM
(
SELECT A.user1, A.user2,
COUNT(*) as share_cnt
```

```
FROM table1
JOIN table2
ON A.user1=B.sender_id AND A.user2=B.recipient_id
WHERE content_type='share'
GROUP BY 1,2

UNION ALL

SELECT A.user1, A.user2,
COUNT(*) as share_cnt
FROM table1
JOIN table2
ON A.user2=B.sender_id AND A.user1=B.recipient_id
WHERE content_type='share'
GROUP BY 1,2
) tb
GROUP BY 1,2
```

# 8. Customers Who Never Order

Suppose that a website contains two tables, the Customers table and the Orders table. Write a SQL query to find all customers who never order anything.

Table: `Customers`.

```
+----+-------+
| Id | Name  |
+----+-------+
| 1  | Joe   |
| 2  | Henry |
| 3  | Sam   |
| 4  | Max   |
+----+-------+
```

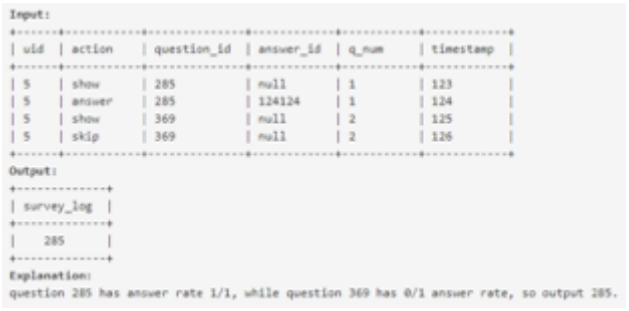
Table: `Orders`.

```
+----+------------+
| Id | CustomerId |
+----+------------+
| 1  | 3          |
| 2  | 1          |
+----+------------+
```

```
SELECT A.Name
FROM Customers AS A
LEFT JOIN Orders AS B
ON A.Id=B.CustomerId
WHERE B.Id is NULL
```

# 9. Answer Rate

table: |user_id|question_id|question_order||action|timestamp| Find the question id with highest answer rate:



```
Input:
+------+-----------+--------------+------------+-----------+------------+
| uid  | action    | question_id  | answer_id  | q_num     | timestamp  |
+------+-----------+--------------+------------+-----------+------------+
| 5    | show      | 285          | null       | 1         | 123        |
| 5    | answer    | 285          | 124124     | 1         | 124        |
| 5    | show      | 369          | null       | 2         | 125        |
| 5    | skip      | 369          | null       | 2         | 126        |
+------+-----------+--------------+------------+-----------+------------+
Output:
+-------------+
| survey_log  |
+-------------+
|    285      |
+-------------+
Explanation:
question 285 has answer rate 1/1, while question 369 has 0/1 answer rate, so output 285.
```

```
SELECT tb.question_id
FROM
(
SELECT question_id,
SUM(IF(action=”answer”, 1, 0)) / COUNT(*) AS rate
FROM table
GROUP BY 1
ORDER BY 2 DESC
LIMIT 1
) tb
```

Follow up:

- How to dynamically change the order of the questions showing to the users to achieve the highest conversion rate (Similar question: We have a table {Survey: saw/skip/answered} along with question_order & question_id. If a new user skips the question with highest & second highest frequently answered question, how should we change the order of questions so we can get more answered question (highest conversion rate)?)
- What should we do to the questions with only a few ‘show’ records (如果imp太少的问题怎么办)
  我答的是加个threshold，舍弃那些出现少于多少次的问题（这里我随便写的10，他没有说什么，不知道对不)
- 第二问是在用户已经回答了某一问题的情况下，如何安排下一问题使conversion rate最高，我这里就按地里讨论的一样说在已经回答了这一问题的用户中，选他们回答的其余问题里回答率最高的一个

```
select question_id
from survey_log
where user_id in
(
select user_id
from survey_log
where question_id = "回答过的那题的ID"
and event = 'answered'
)  tmp
and question_id = "回答过的那题的ID"
group by question_id
order by
sum(case when event = 'answered' then 1 else 0 end)/(case when event = 'img' then1 else 0 end) desc
limit 1
```

# 10. ADs

Q1: Advertiser spend and ROI metrics


Given the following two tables:

- Advertiser info table with columns `advertiser_id`, `ad_id` and `spend`, where spend is what the advertiser paid for this ad.
- Ad info table with columns `ad_id`, `user_id` and `price`, where `price` is how much the `user_id` spend through this ad., assuming all prices > 0.

Questions

- The fraction of advertisers has at least one conversion.
- What metrics would you show to advertisers.

1) The fraction of advertiser has at least 1 conversion?

```
SELECT
SUM(IF(B.ad_id is NULL,0,1))/count(A.ad_id) as rate
FROM adv_info AS A
LEFT JOIN ad_info AS B
ON A.ad_id=B.ad_id
```

2) What metrics would you show to advertisers (其实就是在问ROI), 用SQL实现

```
SELECT advertisers_id,
sum(IF(B.price is NULL,0, B.price)) / sum(A.spend) as ROI
FROM adv_info AS A
LEFT JOIN ad_info AS B
ON A.ad_id=B.ad_id
GROUP BY advertisers_id
```

3) {table1: advertiser_id, ad_id, spent}, {table2: uid, ad_id, purchase}. 求过去三十天advertiser花了多少钱在ad上(per advertiser), distribution是什么, why? 然后by adid求ROI

COST:

```
SELECT A.total,
COUNT(DISTINCT A.advertiser_id) as freq
FROM
(
SELECT advertiser_id,
SUM(spent) as total
FROM table1
WHERE date BETWEEN curdate() AND curdate()-30
GROUP BY 1
)A
```

ROI:

```
SELECT A.ad_id,
SUM(IFNULL(B.purchase, 0)) / (SUM(A.spend)) AS ROI
FROM table1 AS A
LEFT JOIN table2 AS B
ON A.ad_id=B.ad_id
GROUP BY 1
```

Follow up question: In which case ROI is not the best metric?

when advertiser cares less about revenue but more about CTR (eg. marketing campaign), ROI is not the best metric

Q2: 24. unit就是一个他未来会买到的广告的template，一个user可以看同一个unit很多次，也可以看到不同的unit，如果user create an ad了的话就不会再看到了。两个表:
ad4ad: date, user_id, event(impression, click, create_ad), unit_id, cost, spend, ad_id

- ad_id: event是create_ad对应的行才有数值
- event有三个值，其中impression就是FB为user创造的广告space，click代表user点进去了，create_ad代表user点进去之后正式create了ad
- cost就是FB为这个user的这个item创造广告space所花费的cost
- spend就是在user create_ad之后pay给FB的，只有create_ad那一行的spend不是null

users: user_id, country, age

1) last 30 days, by country, total spend (问的是facebook的spend就是表里的cost

```
SELECT A.country, SUM((IFNULL(B.cost,0)) AS num_users
FROM users AS A
LEFT JOIN ad4ad AS B
ON A.user_id = B.user_id
WHERE DATEDIFF(A.date, CURDATE()) <= 30
GROUP BY A.country
```

2) how many impressions before users create an ad, given an unit?

```
SELECT AVG(sub.num_impression) AS avg_imp
FROM
(
SELECT a.user_id, a.unit_id,
SUM(IF(a.event = 'impression', 1, 0) num_impression
FROM ad4ad AS a
JOIN ad4ad AS b
ON a.user_id = b.user_id AND a.unit_id=B.unit_id
WHERE a.date < b.date
AND a.event = 'impression'
AND b.event = 'create_ad'
GROUP BY a.user_id, a.unit_id
) sub
```

Product:

1) metrics to measure the health
2) Which one is the most important to show to CEO? Prodit
3) If one metric goes down, what is the reason?
4) 从Facebook角度来说推出这个feature有什么好坏?

Q3:一个table有三列，user_id，ad_id，time_stamp。问对于每个user-ad pair，找出如果一个用户看了一个广告超过两次，那么最后一次以及倒数第二次看广告之间的时间差。

```
SELECT A.user_id, A.ad_id,
MAX(A.last)-MAX(A.previous_t) as diff
FROM
```

```
(
SELECT t1.user_id, t1.ad_id, t2.timestamp as last
LAG(t2.timestamp) OVER(PARTITION BY t1.user_id, t1.ad_id ORDER BY t2.timestamp) as previous_t

FROM
(
SELECT user_id, ad_id
FROM table
GROUP BY user_id, ad_id
HAVING COUNT(ad_id)>2
) t1

JOIN

(
SELECT *
FROM table
) t2 ON t1.user_id=t2.user_id AND t1.ad_id=t2.ad_id

) A
GROUP BY 1, 2
```

# 11. Comments

Q1: Stroy comments distribution
#Table name: content_actions
- user_id
- content_id
- content_type ('post', 'photo', 'comment') #story: post or photo
- target_id

- If content_type=’comment’, then target_id is content_id(post)
- If content_type=’post’, then target_id is NULL

1) Generate a distribution for the #comments per story.

```
SELECT tb.comments, COUNT(*) as freq
FROM
(
SELECT A.content_id as story,
CASE WHEN b.target_id is not NULL THEN b.cnts
ELSE 0 END AS comments
FROM table AS A

LEFT JOIN

(
SELECT target_id, COUNT(user_id) as cnts
FROM table
WHERE content_type=”comment”
GROUP BY 1
```

```
) B ON A.content_id = B.target_id

WHERE A.content_type in (''post'', ''photo'')
) tb
GROUP BY 1
```

2) Does this account for stories with 0 comments? YES
3) If now content_type becomes (post, video, photo, article), calculate the comment distribution of each content_type

```
SELECT tb.comments, tb.type, COUNT(*) as freq
FROM
(
SELECT A.content_id, A.content_type as type,
CASE WHEN b.target_id is not NULL THEN b.cnts
ELSE 0 END AS comments
FROM table AS A

LEFT JOIN

(
SELECT target_id, COUNT(user_id) as cnts
FROM table
WHERE content_type=”comment”
GROUP BY 1
) B ON A.content_id = B.target_id

WHERE A.content_type in (''post'', “video”, “photo”, “article")
) tb
GROUP BY 1,2
```

4) 如果不看date range, data太大怎么办?
- 我说就看今天的，他问那今天的有什么问题。就是没法capture cumulative num of comments, 只有今天的。
- only use the data for the region/platform we are interested in
- random sampling

5) 你现在有# of comments for a certain post, 你怎么知道这个number是reasonable的。
取些sample看variance, confidence interval

Q2: Comment on post

user_data

| user_id | current_link_id | previous_link_id | data_type |
|---|---|---|---|
| 1 | 2131 | | post |
| 2 | 53123 | | picture |
| 3 | 213 | | movie |
| 6 | 1231 | 2131 | comment |
| 11 | | 53123 | comment |
| 12 | | 53123 | comment |
| 13 | | 53123 | comment |
| 14 | | 53123 | comment |

1) percent of post having at least one comment

```
SELECT
SUM(IF(B.previous_link_id IS NULL,0,1)) / COUNT(A.*) as p_one_comment
FROM
(
```

```
SELECT *
FROM user_data
WHERE data_type=”post”
) A

LEFT JOIN

(
SELECT *
FROM user_data
WHERE data_type=”comment”
) B ON A.current_link_id=B.previous_link_id
```

2) percent of post having at least 5 comments

```
SELECT
SUM(IF(tb.cnt>=5,1,0)) / COUNT(*) AS p_five_comments
FROM
(
SELECT A.current_link_id,
SUM(IF(B.previous_link_id IS NULL,0,1)) as cnt
FROM
(
SELECT *
FROM user_data
WHERE data_type=”post”
) A

LEFT JOIN

(
SELECT *
FROM user_data
WHERE data_type=”comment”
) B ON A.current_link_id=B.previous_link_id
) tb
```

# 12. CTR

TABLE: {time, user_id, app_id, event ('imp' or 'click')}.
现在有一张表，有time, user_id, app, event（impression，click），每个用户在每个app上有一定几率弹出一个窗口填写信息，如果填写了event为click，如果只看见没填写为impression，没看见为空。

1) 求这个功能的click through rate。

```
SELECT SUM(IF(event=’click’, 1, 0) / COUNT(*) AS CTR
FROM table
```

2) 如果CTR>100%是什么原因?

- how do we counted the page view (that is, how to define impression?)

    - This happens because users are able to access your ad in two different ways during a single browsing session. For example, they can open the ad and the sitelink by using a new window or new tab. Whenever a user does this, one ad impression generates two separate clicks, causing the CTR to double. Sometimes it will continue past 200%, resulting in rates as high as 500%.
    - The first click was a quick peek, and then they return with the second click after checking out other alternatives.
    - In some cases, data for your ad campaign comes from a number of unique servers. The data for impressions may come from one server, while the data for clicks is from another.
- how do we counted the unique view (that is, how to define click?)
    - Invalid click like: when a user double-clicks on your Ad; dubious clicks that are used to increase profits for the website hosting your Ad; manual clicks that are intended to increase your advertisement cost.
- how to define the unit: by user/page/cookie? How to combine different device? What time interval we choose to define uv and pv?

3) 如果现在每一个impression可能对应多个click，如何从所有click记录里面选出正确的那个记录来计算Write a query for CTR.
4) what if there is a lot of pollution in this table?

# 12. Message

table: date | user_A | user_B | number_messages
Each row has the number of messages between a unique user pair

1) what can you tell from this table: how many messages between users per day
2) Write a query for the distribution of number of messages for each user. How the distribution is gonna look like and why?

```
SELECT tb.cnt,
COUNT(tb.user_A) as freq
FROM
(
SELECT user_A,
SUM(number_messages) as cnt
FROM table
GROUP BY user_A
) tb
GROUO BY 1
ORDER BY 1
```

3) Write a query find the top partner for each user (most messages)

```
SELECT tb2.user_A, tb2.user B
FROM
(
SELECT tb1.user_A, tb1.user B,
RANK() OVER(PARTITON BY tb1.user_A ORDER BY tb1.total) as rank
FROM
(
SELECT user_A, user_B,
```

```
SUM(number_message) OVER(PARTITION BY user_A, user_B) as total
FROM table
) tb1
) tb2
WHERE tb2.rank=1
```

# 非典型问题

1. An attendance log for every student in a school district {attendance_events : date | student_id | attendance}
A summary table with demographics for each student in the district {all_students : student_id | school_id | grade_level | date_of_birth | hometown}
Using this data, you could answer questions like the following:

1) What percent of students attend school on their birthday?

```
SELECT AVG(A.attendance)
FROM attendance_events AS A
JOIN all_students AS B
ON A.student_id=B.student_id AND A.date=B.date_of_birth
```

2) Which grade level had the largest drop in attendance between yesterday and today?

```
SELECT tb1.grade_level,
MAX(tb1.today-tb2.yesterday) AS drop
FROM
(
SELECT B.grade_level, AVG(A.attendance) as today
FROM attendance_events AS A
JOIN all_students AS B
ON A.student_id=B.student_id
WHERE date=curdate()
GROUP BY 1
) tb1

JOIN

(
SELECT B.grade_level, AVG(A.attendance) as yesterday
FROM attendance_events AS A
JOIN all_students AS B
ON A.student_id=B.student_id
WHERE date=curdate()-1
GROUP BY 1
) tb2 ON tb1.grade_level=tb2.grade_level
```

2. Create daily tracking table of user status
Given a table that each day shows who was active in the system and a table that tracks ongoing user status, write a procedure that will take each day's active table and pass it into the ongoing daily tracking table. Possible states are:

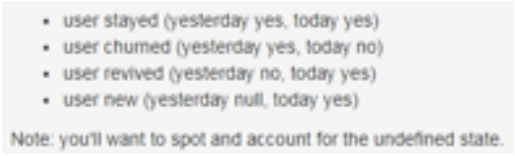



TABLE1: Tracking {user, status}
TABLE2: Day {user}

How do you calculate monthly active users, churned users and resurrected users from a user activity log with userID and DateTime

3. week over week change% (date | user_num, 问如何找到top 100 week over week increase/drop)

```
SELECT tb2.weeknum, ABS(tb2.cnt-tb1.cnt) as change_by_last_week
FROM

(
SELECT tb1.weeknum,
SUM(tb1.user_num) over(partition by tb1.week_num) as cnt
FROM

(
SELECT *,
WEEK(date) as week_num
FROM table
) tb1

) tb2

LEFT JOIN

(
SELECT tb1.weeknum,
SUM(tb1.user_num) over(partition by tb1.week_num) as cnt
FROM

(
SELECT *,
WEEK(date) as week_num
FROM table
) tb1

) tb3 on tb1.weeknum=tb2.weeknum-1

ORDER BY change_by_last_week DESC
LIMIT 100
```

4. There is a table that tracks every time a user turns a feature on or off, with columns user_id, action ("on" or "off), date, and time.

1) How many users turned the feature on today?

```
SELECT COUNT(DISTINCT user_id) as cnt
FROM table
WHERE date=curdate()
AND action=”off”
```

2) How many users have never turned the feature on?

```
SELECT COUNT(DISTINCT user_id)
FROM table
WHERE user_id NOT IN
(
SELECT user_id
FROM table
WHERE action=”on”
)
```

3) In a table that tracks the status of every user every day, how would you add today's data to it?

```
INSERT INTO table_name (column1, column2, column3, ...)
VALUES (value1, value2, value3, ...);
```

If you are adding values for all the columns of the table, you do not need to specify the column names in the SQL query. However, make sure the order of the values is in the same order as the columns in the table. The INSERT INTO syntax would be as follows:

```
INSERT INTO table_name
VALUES (value1, value2, value3, ...);
```

5. table 1：countryid | date | sales id | sales name | amount
table 2：countryid, country
1）问每个地区的总销售金额

```
SELECT country_id,
SUM(amount) as total
FROM table1
GROUP BY 1
```

2）问每个自然月，分地区销售额最高的销售人员id、name、amount，如果有多个，取avg

```
SELECT tb.sales_id, tb.sales_name,
AVG(tb.amount) AS average
FROM
(
select salesid,
salesname,
amount,
dense_rank()over(partition by salesid order by amount desc) as amountRank
from table 1
) tb
```

```
WHERE tb.amountRank=1
GROUP BY 1,2
```

6. given Users {id|country}, Videos{userid|videoid|duration|...}
    1) find total_watch_time for each country

```
SELECT country,
SUM(duration) as total_watch_time
FROM Users AS A
JOIN Videos AS B
ON A.id=B.userid
GROUP BY 1
```

    2) find top three country with highest total watch time

```
SELECT temp.country
FROM
(
SELECT tb.country,
DENSE_RANK() OVER(PARTITION BY tb.country ORDER BY tb.total_watch_time) as rank
FROM
(
SELECT country,
SUM(duration) as total_watch_time
FROM Users AS A
JOIN Videos AS B
ON A.id=B.userid
GROUP BY 1
) tb
) temp
WHERE temp<=3
```

    3) find average of watch time for non-top 3 countries (answer is just one number)

```
SELECT AVG(temp.total_watch_time) as avg
FROM
(
SELECT tb.country, tb.total_watch_time
DENSE_RANK() OVER(PARTITION BY tb.country ORDER BY tb.total_watch_time) as rank
(
SELECT country,
SUM(duration) as total_watch_time
FROM Users AS A
JOIN Videos AS B
ON A.id=B.userid
GROUP BY 1
) tb
) temp
WHERE temp>3
```

7. 三张 tables 分别是 (user_name, sports), (userid, user_name, registration date), (follower id, followee id, following date) 求

1) How many people are following each sports category?

```
SELECT t1.sports, COUNT(DISTINCT t3.follower_id) as cnt
FROM t1
JOIN t2 ON t1.user_name=t2.user_name
JOIN t3 ON t2.user_id=t3.follwer_id
GROUP BY 1
```

2) How many basketball players are following football players?

```
SELECT COUNT(DISTINCT A.user_id) as cnt
FROM
(
SELECT t2.user_id
FROM t1
JOIN t2 ON t1.user_name=t2.user_name
WHERE t1.sports=”basketball”
) A
JOIN
(
SELECT t3. follower_id, t3.followee_id
FROM t1
JOIN t2 ON t1.user_name=t2.user_name
JOIN t3 ON t2.user_id=t3.follwee_id
WHERE t1.sport=’football’
) B ON A.user_id= follower_id
```

# FB DS Interview

## How to prepare:

- Product design: Use the product design framework CIRCLES Method to explore possible personas and articulate the use cases.
- Execution: get things done and make critical decisions. AARM Method, Google HEART framework for User Experience
- leadership + drive : What is your biggest weakness?

### High-Level Talks

Alex Schultz — How to Get Users and Grow Youtube
Chamath Palihapitiya — how we put Facebook on the path to 1 billion users Youtube
Alistair Croll — Lean Analytics: Using data to build a better startup faster Youtube

FB Official Numbers: Link
Instagram stats: Link Link2

### Examples: StellarPeers

How would you measure the success of Facebook Stories? Link
How would you find the cause of a 15% drop in Facebook Groups usage? Link

### 经验总结帖:

如何准备Data science analytics interview, case study详解 by Alice007 Link
总结自己如何cracking the Data Challenge by Alice007 Link
product sense的经验+鸡汤 Link
非典型性面经（Facebook, LinkedIn, Pinterest） Link

### 地里相关面经

**1.**Link Link2 Link3 Link4

SQL – Composer
table composer，3 columns: userid l event l date,
event 包括 enter/post/cancel (enter就是开始在composer里面写内容，cancel就是开始编辑但是没有post而是终止了)
(1) what is the post success rate for each day in the last week?

```
SELECT
        date,
        ROUND(IFNULL(SUM(CASE WHEN event = 'post' THEN 1 ELSE 0 END)/
                                SUM(CASE WHEN event = 'enter' THEN 1 ELSE 0 END), 0), 2) AS
success_rate
FROM
        composer
WHERE
        DATEDIFF(CURDATE(), date) <= 7
GROUP BY 1;
```

(2) 在第一题的基础上，又给了一个table：user，4 columns: userid | date | country | dau_flag{0, 1}。其中dau_flag表示daily active or not
what is the average number of post per daily active user by country today?

```
SELECT
    country,
    ROUND(IFNULL(SUM(CASE WHEN C.event = 'post' THEN 1 ELSE 0 END)/
COUNT(DISTINCT U.userid), 0), 2) AS avg_posts
FROM
    user U
    LEFT JOIN
    composer C
    ON U.userid = C.userid AND U.date = C.date
WHERE
    date = CURDATE() AND U.dau_flag = 1
GROUP BY 1;
```

Product Sense
Given a tracking metric: Avg_ActiveUser_post -> The average active user post number per day:
We have a drop from 05/01/2018 to 06/01/2018 below：
05/01/2018 Avg_ActiveUser_post 3.0
06/01/2018 Avg_ActiveUser_post 2.5
Question: What is the reasons/factors you would think that cause this drop ?
Open. As long as it makes sense.

1.问的是上面的metric - average number of post per daily active user 突然从3下降到2.5，有哪些可能的原因，并且解释每个原因。

(1) ask for clarifications: problem with data collection? one-time event or progressively? seasonality? platform? region? special events?
(2) about the metric: numerator — any change in total # of posts from all DAU per day? privacy concern? a similar feature/product?
denominator — any change in # of DAU? new users are not less willing to reveal themselves online?

好像还问了个问题，是怎么样确定一个新的change是好是坏之类的，有哪些metric可以帮助measure。
What is the goal of this change? profit or engagement?
The goal of Facebook is to increase user engagement and retention, a new feature which helps to achieve this goal is good. To measure engagement/retention, use metrics like # of active users daily/monthly, fraction of users who used this new feature. avg # of posts per DAU, time spent

我问了是worldwide么，然后check了有没有突发情况或者seasonality，然后说大家either total number of posts下降了or dau增多了但是增加的dau发帖子不活跃，
follow up问怎么确定dau活跃程度，我说bucket by time on fb
time spent, avg # of posts/comments/likes

2. fb app现在想改版成ins app那种在页面最下方有一个发帖按钮的interface，问怎么设计a/b

testing

## 2.Link Link2

SQL — message
Table: date, timestamp, send_id, receive_id
Question: What's the fraction of users communicating to > 5 users in a day?

```
SELECT
        date,
        ROUND(SUM(CASE WHEN num_contacts >= 5 THEN 1 ELSE 0)/COUNT(userid), 2) AS
fraction
FROM
        (SELECT date, userid, COUNT(DISTINCT userid2) AS num_contacts
        FROM
        (SELECT date, send_id AS userid, receive_id AS userid2
         FROM message
        UNION ALL
        SELECT date, receive_id, send_id
        FROM message
        ) T1
        GROUP BY 1, 2
        ) T2
GROUP BY 1;
```

Product — Best friend?
# of interactions (likes, comments, post on timeline), # of photo tags, gifts, messenger data, demographical data

## 3.Link Link2

SQL — sms_message (fb to users)

| date | country | cell_numer | carrier | type |
|---|---|---|---|---|
| 2018-12-06 | US | xxxxxxxxxx | verizon | confirmation (ask user to confirm) |
| 2018-12-05 | UK | xxxxxxxxxx | t-mobile | notification |

confirmation (users confirmed their phone number)

| date | cell_number |
|---|---|

(User can only confirm during the same day FB sent the confirmation message)

1. yesterday how many confirmation texts by country.

```
SELECT
        country, COUNT(*)
FROM
        sms_message
WHERE
        type = 'confirmation' AND DATEDIFF(CURDATE(), date) = 1
GROUP BY 1
ORDER BY 2 DESC;
```

2. Number of users who received notification every single day during the last 7 days.

```
SELECT
        date, COUNT(DISTINCT cell_number) AS num_users
FROM
        sms_message
WHERE
        DATEDIFF(CURDATE(), date) <= 7  AND type = 'notification'
GROUP BY 1;


SELECT
        COUNT(*)
FROM
        sms_message
WHERE type DATEDIFF(CURDATE(), date) <= 7  AND type = 'notification'
GROUP BY cell_number
HAVING COUNT(DISTINCT date) = 7


3. 过去30天的 confirmation rate
SELECT
        T1.date, ROUND(IFNULL(num_confirmations/num_send, 0), 2) AS confirmation_rate
FROM
        (SELECT date, COUNT(*) AS num_confimations
         FROM confirmation
         WHERE DATEDIFF(CURDATE(), date) <= 30
         GROUP BY 1;
        ) T1
        JOIN
        (SELECT date, SUM(CASE WHEN type = 'confirmation' THEN 1 ELSE 0 END) AS
num_send
         FROM sms_message
        WHERE DATEDIFF(CURDATE(), date) <= 30
         GROUP BY 1
        ) T2
        ON T1.date = T2.date
GROUP BY 1;
```

如果有一个简化版的FB，只有简单的文字post和comment且comment 没有nested这个功能（只能一条一条）。 问如何判断一个post的comment包含conversation。有什么metrics可以监测。comments的话可以看有没有几个user back and forth的pattern

然后问了几道probability，FB ads 有lazy reviewer 和common reviewer, 一个reviewer是common的概率是0.8，是lazy概率是0.2。common给好评概率是0.6,差评0.4。lazy全给好评。问1.）一条ads 是好评的概率，2.) 100个ads里number of 好评的expectation。3.）有五个ads都是好评，是lazy的概率
(1)P(good) = 0.8 * 0.6 + 0.2 = 0.68  (2) E = 100 * 0.68 = 68 (3) P(lazy| 5 good) = 0.2/(0.6^5 * 0.8 + 0.2) = 0.76

**probability**:
two approaches:
a. 5% chance to be an ad per post.
b. every 20 post must have an ad in it.
1. compute each expected value and variance for number of ads in 100 posts.
expected = 5 for both, var_a = 100 * 0.05 * 0.95 = 4.75, var_b = 0

2. probability of getting more than 10 ads in 100 posts with approach a.

我是这么想的:
approach a应该符合二项分布，p=0.05, q = 1-p = 0.95.
如果用二项分布解的话:
p(more than 10 ads) = 1 - p(less than or equal 10 ads)

p(less than or equal 10 ads) = p(ads = 0) + p(ads=1) + ... + p(ads=10)
等式右边的每一项可以用二项分布的概率密度函数解。。

但是这么解会比较耗时间。可以考虑用正态分布去估计（因为我们看到样本数目比较大np>=5且nq>=5）:
此处 mu = 5, var = 100*0.05*0.95 = 4.75, sigma = sqrt(4.75) = 2.18
**Z = (x - mu)/sigma = (10 - 5)/2.18 = 2.29**
如果没有Z表，可以这么估计:(我不太确定这么估计是否会让面试官满意，但是应该比没有估计好吧。。)
我们很熟悉的单尾Z<1.96的概率是0.975, 所以p(Z>1.96) = 0.025
所以p(Z>2.29) < p(Z>1.96) = 0.025
解答:
100个posts中有超过10个广告的概率不超过2.5%，具体数字根据查表得到为1.1%。
根据二项分布的解法，用Excel算了一下，结果是:1.1472%

3. expected number of seeing back-to-back ads in 100 posts with two approaches.

**product**:
why facebook require users to register accounts with phone number or email address confirmation? What pros and cons each have?
(1) recover password/account
(2) send notifications
(3) import contacts and find friends to improve engagement
(4) detect duplicated accounts
(5) privacy
(6) phone message may not be free, email requires internet connection

**product**: Link
instagram now have feature let users to switch accounts with one button
1. how to identify multiple accounts belonging to the same user?
(1) phone number/email address (2) user_name (3) following/follower intersection (4)

device id

2. total timespent flats and number of accounts increases. What are you hypothesis about why this happened? what data you need and how to testify your hypothesis? how do you determine if this feature is successful launch?
我的回答是可能有novelty effect，一开始推出这个feature大家觉得有趣就去多建了几个账号但是the way ppl interact with Instagram并没有变化所以avg time spent没有变。是否要 launch就要结合opportunity size以及你想的一些metrics来决定是否有practical impact，impact有多大来决定。

## 4.Link

SQL — ad4ad：
两个表
ad4ad: date, user_id, event(impression, click, create_ad), unit_id, cost, spend，记得还有一列，可能是ad_id，event是create_ad对应的行才有数值
users: user_id, country, age

ad4ad的背景，简单来说unit就是一个他未来会买到的广告的template，一个user可以看同一个unit很多次，也可以看到不同的unit，如果user create an ad了的话就不会再看到了。我补充一下这个ad4ad，fb想让一些通过fb宣传自己产品的隐性广告购买者也成为paid users。方法就是给他们提供一些广告的template，这些users就会在自己的newsfeed不断看到这些为他们设计的template，每次看到都算作一次impression，如果进一步点击了就算一次click，最后购买了的话就会create an ad。这样算一次成功的转化。

1. last 30 days, by country, total spend (问的是facebook的spend就是表里的cost） of the product

```
SELECT
        country, SUM(IFNULL(spend, 0)) AS total_spend
FROM
        users U
        LEFT JOIN
        ad4ad A
        ON A.user_id = U.user_id
WHERE DATEDIFF(CURDATE(), date) <= 30
GROUP BY country;
```

2. how many impressions before users create an ad given an unit?

```
SELECT
        t1.user_id, t1.unit_id, SUM(CASE WHEN event = 'impression' THEN 1 ELSE 0 END) AS num_impression
FROM
        (SELECT DISTINCT user_id, unit_id
         FROM ad4ad
         WHERE event = 'create_ad'
        ) t1
        LEFT JOIN
        ad4ad
```

```
        ON t1.user_id = ad4ad.userid AND t1.unit_id = ad4ad.unit_id
GROUP BY 1, 2;
```

3.avg number of impressions per user per item before user creates ad

```
SELECT
        unit_id, SUM(num_impression)/COUNT(DISTINCT user_id)
FROM
        (SELECT t1.user_id, t1.unit_id, SUM(CASE WHEN event = 'impression' THEN 1 ELSE 0
END) AS num_impression
         FROM
                (SELECT user_id, unit_id
                FROM ad4ad
                 WHERE event = 'create_ad'
                ) t1
                LEFT JOIN
                ad4ad
                ON t1.user_id = ad4ad.userid AND t1.unit_id = ad4ad.unit_id
        GROUP BY 1, 2;
        ) T2
GROUP BY 1;
```

product1：list few metrics related to ad4ad, why these metrics. If one metric goes down, what is the reason?
再说一下关于product的问题，如果一个metric go down了，大家不仅仅要break down the metric from user perspective(e.g. country, device, etc), 同时要想到这个metric是怎么算的，denominator和numerator是什么
metric下降可能是numerator变小了，可能是什么原因导致；然后denominator增加有可能是什么原因导致。还有如果别的创造广告的途径的收入下降了，对ad4ad好不好呢，可能会有什么影响呢等等，总之问的很细

avg就说不够robust to outliers，可以考虑用median,
metric 还可以提到revenue per targeted user = total ads revenue generated by ad4ad / total target ad4ad users

Product2:
1. metrics to measure the health:
        基于这个产品的feature，转换率还有profit是比较重要的。# or % of  users who used this feature per month; # of ads posted per month; overall, profit through this feature per month?
2. Which one is the most important to show to CEO? Profit
3. 你的角度都不错，也给了我一些启发。
我建议从profit的计算方法开始展开，也即从revenue和cost的两个角度出发。revenue低，可能是定价低，可以提高价格。但与此同时，或许购买人数会变小。这里我们需要做一个optimization。如果是cost的问题，有可能是购买率低，每次成功的转换都需要太多impression，说明我们推荐的不够好，不是target users。
4. 从Facebook角度来说推出这个feature有什么好坏?

广告太多也会影响其他用户的使用体验等等。

Link2
Product:
你会用什么metrics来衡量这个feature？如果有个metrics下降了10%， 你会怎么做，你可以看到fb任何数据，你会怎么来做（大概意思就是你怎么segment去看问题出现在哪里）

因为我觉得我product回答的都还可以的， 但是被通知要加试一轮product analysis，所以做出了一下总结，希望后面的人不要犯同样的错误！
附上总结，1.在回答product的问题的时候，不要emmmm。。。面试官会以为你被难住了，你不会而且没自信，其实我这只是我的个人习惯我只是想表示我在思考，其实完全可以说： please kindly allow a moment fororganizing my thoughts.
2. 在说metrics的时候不要想到哪个metrics就甩过去，因为你可能会被follow up很深。这次就是被问了很深，幸好自己比较了解我自己选择的几个metrics。Metrics哪里好哪里不好？要有准备会问到。还会被问到哪个metrics更好。如果当时脑子卡了一下，可以问一下这个公司这个产品当前的目标，对你进一步选择metrics 有帮助，
3. metrics下降了这种问题，不要只想着denominator，numerator， 也要顺着产品去想想问题出现在哪里
4. 最后的忠告是千万别犹豫！！！如果说 当你想选择table context里面的segment，千万别犹豫，说出来，是对的，没必要一定要自己想出来个什么来，脸皮不要薄。想到啥说啥，千万别怕错，你想说的很可能是对的！！

## 5.Link Link2 Link3 Link4 Link5

SQL — SPAM
-- Q1: how many posts were reported yesterday for each report Reason?
-- Table: user_actions
-- ds(date, String) l user_id l post_id l action ('view','like','reaction','comment','report','reshare') l extra (extra reason for the action, e.g. 'love','spam','nudity')

```
SELECT
        extra AS reason,
        COUNT(IDSTINCT postid)
FROM
        user_actions
WHERE action = ‘report’ AND DATEDIFF(CURDATE(), date) = 1
GROUP BY 1;
```

-- Q2: introduce a new table: reviewer_removals, please calculate what percent of daily content that users view on FB is actually spam?
--no need to consider if the removal happen at the same post date or not.
-- ds(date, String) l reviewer_id lpost_id

```
SELECT
        U.date,
        U.user_id,
        COUNT(reviewer_id)/COUNT(DISTINCT U.post_id)AS percentage
FROM
```

```
        user_actions U
        LEFT JOIN
        reviewer_removlas R
        ON U.post_id = R.post_id
GROUP BY 1, 2;
```

Q3: How to find the user who abuses this spam system?
第三问我是output了一个table：第一列是user id，第二列是这个user report了多少个post是spam，第三列是这个user report spam的post中到底是有多少是真的spam，所以就是left join了一下。面试官也说ok能够提供想要的结果。

```
SELECT
        user_id,
        COUNT(reviewer_id)/COUNT(DISTINCT U.post_id) AS fraction_spam
FROM
        user_actions U
        LEFT JOIN
        reviewer_removals R
        ON U.post_id = R.post_id
WHERE
        U.action = 'report'
        AND DATEDIFF(CURDATE(), date) <= 30
GROUP BY 1;
```

Product:

- How would you test if this filter works?
  - num  of spam reported every day
  - num of shares/comments/likes
  - time spent
  - # of active user per day, week, month
- If we experiment, how would you conduct it?
  - A/B testing
- How to select a sample group
  - Random to avoid bias
- How many people would you select for your sample group
  - Use formula for n (minimum sample size)
- After getting results from A/B testing, what to do next?
  - T-test on metrics to see if there's a difference
- What's a t-test? What's t-score? What's P-value? Explain p-value to someone who doesn't know stats.
- Let's say the filter worked but revenue went down, what would be your hypothesis?
  - Revenue comes from click of ads, # users * CTR * price/click, # users and price/click are the same, then CTR gets smaller, people spend more on good contents
  - Short term vs long term, in the long run it's good for the product
- Given revenue decrease, how would you make recommendations? (doesn't have to be yes or no answer)
  - Short term vs. long term: how much does revenue drops? User experience vs. revenue. Short term revenue drop vs. long term brand perception and long term revenue gain.

PRODUCT:
Q3: Facebook has decided to be proactive about SPAM, instead of merely reactive. We decide to address the SPAM problem through a Machine Learning solution predicting whether a given post is Indeed SPAM. We want to use the predictions in order to downrank/deprioritize suspected SPAM from news feed.

Q1. Facebook用machine learning 建了一个model来rank content以达到filter spam的目的，需要关注什么metrics来评价这个model.

Q2. 在用ab testing的时候发现用了新的spam model之后revenue下降了。面试官确定了首先这个model不会touch到ads，就是说ads不会被filter out。并且DAU/WAU/MAU和time spend没有变化，也就是说user方面没有变化。那么可能的原因是什么。
然后我问面试官revenue主要来自什么，面试官说是click ads。我说那么ads click的revenue主要可以break down成#user x CTR/CTP x price/click. 这个情况下只可能有变化的是CTR，也就是说因为用来新的model以后，这个平台的整体content质量更高了，那么user就更喜欢花更多时间去explore这些content，那么点击广告的时间就相对来说哦变少了，revenue也下降了。面试官说是这样的，采用新的model之后用户可能会花更多时间去看video之类的，那么用在ads上的时间就变少了。

#DAU x CTR/CTP x price/click x # ads per user

题目背景介绍很长，大意就是以前spam都是人工看，费时费钱，现在整出来一个ML算法来看。然后筛掉这些。结果发现其实筛掉的不一定对。所以改成了把挑出来的downrank到最下面去。问如何evaluate这个。

楼主一开始有点懵，不知道到底要evaluate算法，还是evaluate downrank这个事情，是和最开始的人工筛选比较还是和之前直接全筛掉比较。

考到的考点有:
1) 哪些metrics? --我选了DAU和avg time spent 衡量engagement。最后的最后烙印告诉我其实他想要的是衡量算法，其实最直观的应该是spam reported before vs after。
2) 样本是random挑选么？ -- 我回答应该从至少report过一次的人里挑，因为一般人可能不举报，说明不敏感。
3) ML这个模怎么建？吓了一跳考起来ML了，幸亏不是很细致，BS了一段
4) spam降了发现engagement也下来了怎么回事？--我依稀记得地里说过distribution，但是当时没细想，临时一问想不起来distribution如何make sense，憋了半天说了一个可能spam的东西有人喜欢看，这些人很生气。烙印又问还可能是什么呢？这下憋不出来了，冷场更多。。。最后烙印放弃换下一题。
5）一个升一个降如何衡量效果。--我说weighted avg来看

Link
**产品部分，**Facebook 做了一个算法来down rank SPAM post。

**1.你会用什么 metrics evaluate the performance of the algorithm?**
这里我分了product和user两个维度来答，首先product方面，可以用spam rate；用户方面，可以用DAU/MAU，以及engagement of posts( likes, comments, shares)。

最后站在公司角度长期来看会影响Revenue，（这里是受面经的影响，直接把后面第三题会问的revenue说了出来....其实这两者看上去是没有直接关系的）面试官这里follow up，让我解释为什么spam filter会影响公司Revenue，因，我有一点点慌，但是还好临时圆了场，说spam 太多会降低用户的使用体验，长期来看也许会流失用户，从而影响公司的revenue。面试官表示make sense。
这里只要能自圆其说应该就问题不大。

**2. how do you know this happens is due to the use of the new algorithm? how do you form your control/experiment group?**
讲一下A/B test的基本流程就好
分组最重要的一点就是Randomization

**3. ads revenue下降了，原因是什么?**
**首先确认没有seasonality / geography 的influence**
然后确定了这个filter不会touch到ads，就是说ads不会被filter out。并且DAU/WAU/MAU和time spend没有变化，也就是说user方面没有变化。
之后 revenue breakdown = #user * price/view * average view （注意这里不是CTR)

if price too low： 面试官说这是ads组的事情，不需要我们来考虑

那就只能是average view下降，面试官让我继续解释，为什么spam filter 会导 ads view下降。这时候我由于受之前面经影响很深，思维固化了，直接说出因为spam filter，这个平台的整体content质量更高了，那么user就更喜欢花更多时间去explore这些content（比如video, photos, posts from friends)，那么看广告的时间就少了。

面试官表示ads view下降 和 time spend无关，这时候我又再次有点慌...后来沉默了一会儿，面试官掏出手机打开insta，给我演示了用户是如何看到广告的，突然我就开窍了，是因为ads是穿插在post里的，如果spam 变少了，user更容易看到他们想看的content，就会减少screen scoll length，相当于看到广告的可能性降低。面试官说make sense。

**4. 这套算法上线后，Spam Rate下降，ads revenue也下降，我们如何决策是否继续这套方案?**
这里由于算法已经上线，所以是一道和A/B test无关的题。

最后整个面试过程结束只花了30分钟，面试官就笑了，说还有十五分钟时间我们要怎么办呢==，我就只能闲扯问了几个fb相关的问题拖到了40分钟。
面试官说fb里一切都是以product为先，无论是engineer, desiner, 还是data scientist都是以产品为单位来分成组，所以每个岗位都需要懂产品，这一点我非常赞同。

## 6.Link
SQL — given table with: (user, group, time, displays, clicks) for a payment page.
1） # of clicks and displays in given day

2) click through rate,
3) click through rate for each group,

4) group 1 click rate: 10%, group 2: 15%, think about possible differences, group 3 click rate 150%, think about possible reason,
5) how to identify click but not displays

Stat
1） 1000 people, each time select 10 (w/o replacement) 问每个人on average多少次会被抽到，2) 1000 people, each time select 10, (w/ replacement), 问每个人on average多少次会被第一次抽到，3）画distribution of page shared (#users vs # page share). 标mean, median, p1, p99, 问根据经验mean是多少，median是多少，取出day 1所有page share=2的population, 问trends for day 2-30, 同样取出day 1 所有page share=5的population, 问同样的trend. 那个方差大，是什么分布?
Explanation: Link



product
问父母加入脸书对子女的影响，1）如何判断父母是否加入，2）直觉上有和影响？3）如何分析数据，4） 如何test
This way works to show correlation between two, but cannot show causal relationship. So you will need to further test this. Couple different ways, 1) design some test to randomize A/B. For example, randomly select from those their parents send them invitation. Just don't show them. (of cause create some customer experience problem), 2) design some treatment. For example, if you think the impact is positive, you send a notification to group A asking them to invite their parents. B group is control. If you think the impact is negative, for those parents already connect with them, Group A introduce a "enhanced privacy setting" to opt them out of their parents post, group B for control. cons: the impact is treatment+behavior

## 7.Link Link2

SQL — table: ad_account,date, spend, status (open,close,fraud)
1.求在active accout里是fraud的概率（active means spend>0）

```
SELECT
    SUM(CASE WHEN status = ‘fraud’ THEN 1 ELSE 0 END)/SUM(CASE WHEN spend > 0 THEN 1 ELSE 0 END) AS prob_fraud
FROM
    table;
```

2.求有多少account是今天被label成fraud

```
SELECT
    COUNT(DISTINCT ad_account)
FROM
    table
WHERE
    date = CURDATE() AND status = ‘fraud’;
```

3.如果给被report为fraud的账户申诉的机会，有什么financial benefit
(1) second chance, those who post the ads will like it, user satisfaction
(2) some user may abuse the report mechanism, they report every ad they see. This will definitely stop some potential advertisers

Link2
如果说要准备这类面试的话，我建议你把FB所有产品都点一遍，每一个button每一个feature都在自己脑子里面过一遍，考虑这样几个问题："这个feature到底有什么用？本质是什么?为什么FB非要加这个feature？它到底有什么独特之处？假设它没有launch，我在launch之前要考虑哪些因素，怎么去衡量成功？如何去做test？这样做是否有局限？怎么改进？"我自己就是这样把FB所有产品走过一遍的。
如果有一起准备的小伙伴的话，可以互相challenge。
面试时候的话，尽可能把这个当作一个交流，优先梳理框架，在白板上整理出来，而不是一上来直接和贯口一样把metrics都倒出来。

Group: Active Group%, Time Spent on Group/Total Time Spent on Facebook, New Member Growth Rate
Best Friend: About Similarity, %Post/Photo Tagged Together, Messanger records
NewsFeed: Favor towards certain post type? Load Speed? Text font? Internet Quality and Coverage? Device Type?

## 8.Link Link2 Link3
SQL — comment distribution
有content_id, content_type (comment/ post), target_id。如果是comment，target_id就是post的id，如果是post则target_id为NULL。求comment distribution。

```
SELECT
        num_comments, COUNT(post_id) AS num_posts
FROM
        (SELECT target_id AS post_id, COUNT(*) AS num_comments
         FROM table
         WHERE content_type = 'comment'
         GROUP BY 1
        ) T1
GROUP BY 1;
```

然后加问：如果现在content_type变成post, video, photo, article，要求计算每一个content type的comment distribution。

```
SELECT
        T1.content_type, T2.num_comments, COUNT(T2.post_id) AS num_posts
FROM
        (SELECT content_id, content_type
         FROM table
         WHERE content_type != 'comment'
        ) T1
        LEFT JOIN
```

```
        (SELECT target_id AS post_id, COUNT(*) AS num_comments
         FROM table
         WHERE content_type = 'comment'
         GROUP BY 1
        ) T2
        ON T1.content_id = T2.post_id
GROUP BY 1, 2;
```

SQL，考的是post type distribution那道题:You have one table named content_action which has 5 fields:Date,
User_id (content_creator_id),
Content_id (this is the primary key),
Content_type (with 4 types: status_update, photo, video, comment),
Target_id (it's the original content_id associated with the comment, if the content type is not comment, this will be null)

Question:
1.find the distribution of stories（photo+video） based on comment count?
2.what if content_type becomes (comment/ post), what is the distribution of comments?
3.Now what if content_type becomes {comment, post, video, photo, article}，what is the comment distribution for each content type?

Distribution
想跟大家详细讨论一下科技公司里面特别喜欢问的distribution问题。出题形式一般是，问一个特定的random variable符合什么样的distribution。比较常见的几种题型如下：
1. What's the distribution of the comments per post?
2. What's the distribution of comment length?
3. What's the distirubtion of page shared per person?
在onsite的时候，还有公司喜欢追问，if we take all people that share 2 page in day1, what can be their distribution of page share in day2?

这种问题我翻找了以下，对于1,2,3，常见答案一般喜欢说log normal 或者poisson distribution。原因是因为认为 X=0的概率最大，且容易出现长尾效应，我觉得这种答案是有道理的，但是可能在面试中不是特别足够，所以专开一贴进行讨论。

Firstly,我觉得poisson distribution不是特别合适。传统意义上，poisson distribution一般用来衡量固定时间内，event出现k的概率，event一般为binary（1l0）。本质上poisson distribution 是一种bionomial的特殊形式，在n趋近于无穷，但成功次数的expection lambda相对固定的时候使用。相对应的推导可以看这里，https://medium.com/@andrew.chamberlain/deriving-the-poisson-distribution-from-the-binomial-distribution-840cc1668239 。在这种情况下，其实情况1是比较合适的，我们可以将每一个post可能有的comment为n，when a user leave a comment, the event = 1 other wise 0. 潜在中的话，n的确趋近于无穷。但是这种解释有一个致命的问题，就是每一个post可以产生的post的expectation一定不相同，大v转发的动辄过万，小透明的肯定没人理睬，所以poisson 分布里每次抽样的，得到的expectation都是lambda这个假设被严重违背了。个人觉得，如果是求daily active user/daily likes之类的distribution会更加合适。

排除掉了poisson我觉得log normal其实是一种比较合适的估计。这里我看了两篇论文：https://

www.sciencedirect.com/science/article/pii/S1877050914005006 和 https://epjdatascience.springeropen.com/articles/10.1140/epjds14，一个是关于distibution of the reweets per tweet， 一个是distribution of comment length. log normal 的论证基础是，有一个变量X服从normal distribution，而变量Y = exp（X)，因而Y服从log normal（即log（Y)服从normal distribution）。在tweet的论文中，作者提到了power of law，大概意思是如果一篇文章本来很受欢迎，那么更多人会转发它,造成exponential的效果，如果我们认为大家会转发文章的这件事情本身-r服从normal distribution, 由于叠加效应,最终这篇文章的转发量应该是(1+r)^k, r如果大于0，将呈现几何增长，（1+r）^k在数学上，可以用exp(rk)表示，所以最后的转发数应该服从log normal。我觉得这种说法还是比较有道理。对于distribution of commen length,第二篇论文提出了一种看法，即人们comment发长的意愿是服从normal distribution的，但是由于心理学上的一种效应，人体的感官对comment的长度感应不敏感，当人们的感受成proportional increase 时，实际长度成exponential增长，举个实际例子：大部分人觉得comment打一个字，和打10个字，感觉没有什么不同，都很短，但是实际上comment length却增长了10倍。Research还做了两个实验，想要证明这个观点。如果上述关系成立那么大家对长度的感官服从normal distirbution , L(length of comments)确实就应该服从log normal。

这两个说法我觉得都有一定的根据，欢迎大家讨论与补充，面试中也可以了解更多，发挥更好。

最后对于if we take all people that share 2 page in day1, what can be their distribution of page share in day2?我认为应该是符合normal distribution的，在day1所有share 两个page的population中，在接下来的时间内，the proportion of share less should be symmetric to sharing more.我暂时没有想出比较solid的stats理由，希望大家能对此提出自己的意见和看法。

## 9.Link Link2 Link3 Link4 Link5

SQL — marketplace user log
Session table
Date I sessionid I userid I action (enter/click/send/exit)
Time table （sessionid都是unique的）
Date I Sessionid I time_spent (s)

Q1: Average sessions/user per day within the last 30 days

```
SELECT
    date, COUNT(DISTINCT sessionid)/COUNT(DISTINCT userid) AS avg_num
FROM
    session
WHERE DATEDIFF(date, CURDATE()) <= 30
GROUP BY date;
```

Q2: Time distribution of each user, 先问我大概会是怎么样的一个分部 — exponential distribution ? Not clear

```
SELECT
    total_time, COUNT(DISTINCT userid)
FROM
    (SELECT
        userid, SUM(IFNULL(time_spent, 0)) AS total_time
    FROM
```

```
            session S
            JOIN
            time T
            ON S.sessionid = T.sessionid AND S.date = T.date
    GROUP BY 1) temp
GROUP BY 1
ORDER BY 1;
```

Product — Marketplace
Q1: launch a call-to-action button “Sell Your Product” on the top banner, what's the reason behind this launch?
我说答案可能有两个方向，
第一个是我们的确发现了用户有这个需求，我们通过观察他们的clickstream和timespent是有可能得出他们需要一个button去引导他们完成selling post，所以我们设计了这个button
第二个我说这算是e-commerce行业的一个常识，用户不会去做他看不到的事情，所以我们需要引导用户，这样有利于我们提高CTR和转化率

Q2: How do you evaluate the impact and make sure this button is actually working?
这里个人认为很明显期待你答A/B testing
然后他问了我key metrics怎么设计，control 和test group都是什么；这里注意key metric不能是CTR，因为control group的设计里根本没有这个button，所以CTR会有bias，
我选择的是conversion rate，# of posts of selling product/total session in the test period, 然后他问我为什么分母是这个，我说因为那个时间段里，根据第一题SQL，用户一个session里会做很多事，其中之一就是post，所以这么算我觉得比较合理，他说好的，make sense。

Q3: 现在ran了A/B testing，发现CR反而降低了，问我为什么
我先确定了有没有external factor，seasonality，data collection error； 他说没有
然后我就回到公式本身，我说那就可能是两种情况，# of post 降低了，或者session数量增加了，session增加的原因可能是最近我们做了什么promotion，或者临近毕业季，大家都想卖东西

Q4: 他在这打断我，说ok，就是# of post降低了，问可能的原因
我说那应该是用户的行为有所变化，但不一定是坏事，我们需要先看一下所有用户的time spend有没有增加，我们的 # of transaction有没有增加，买家端的用户变化，以及我们所处的时间段，因为毕业季可能卖家多，开学季买家多之类的；
其次，那就是比较糟糕的情况，就是我们确实流失了一些user engagement，可能他们确实churn to ebay etc。 他说确实是，可能是买家比较多，用户在实验之前po的东西都被买了，现在这个时间段并不是高峰

Q5: 现在我们想做一个recommendation algorithm （就是类似于淘宝的猜你喜欢），你会怎么去设计这个功能
1. item based
我们可以根据每一个用户的购买记录去给他推类似的商品，但是这个方向有一个弊端，就是marketplace不像传统电商平台，这里是个C2C的，所以可能他的购买记录不是很丰富，其次，用户可能不想一直购买类似的商品，可能会引起厌烦
2. user based
这个方向就可以很好的弥补第一个direction的弊端，因为每一个user都是facebook user，我们可以利用他们的network，去找到跟他们类似的user，甚至跟他们在一个group里的人，去给他们推他的朋友买过的东西，这里的similarity是可以用algorithm模拟出来的。

然后他的反馈比较positive，他说是的，这个make sense

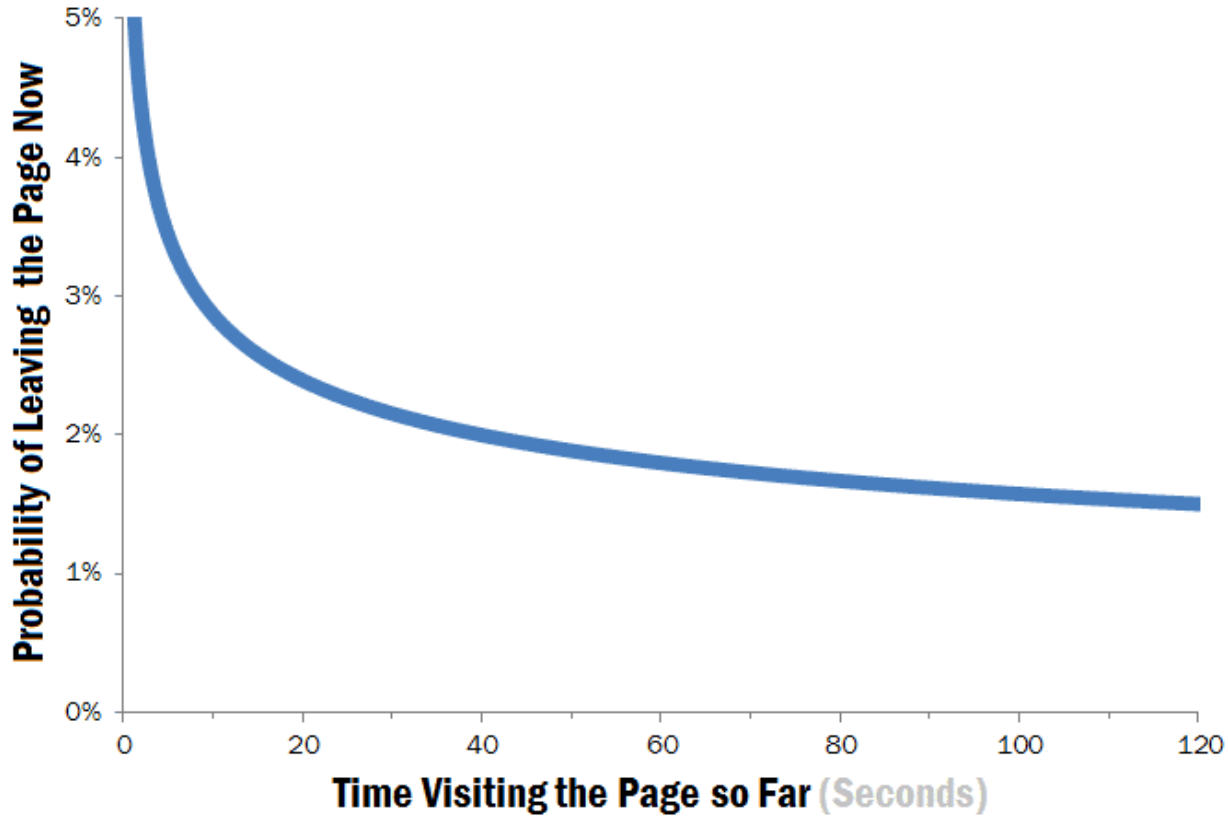
**Tips:**
因为我个人很喜欢facebook，所以准备了很多细节上的小问题
1. 不要小瞧面试开始前的chitchat，那也是你可以抓住机会differentiate自己的地方，一定要比较契合公司文化，建议大家去看看这个链接里的几个重要的value，试着结合自己的经验说说
https://www.facebook.com/careers/

2. SQL的格式要整齐，显得你比较严谨

3. 个人认为这个岗位还是更偏向于product，建议多去读读博客，看看关于key metrics的insights，当然地里的面经也很重要

4. 可以适当的对interviewer下下功夫，我在领英上发现我的interviwer之前创过业，所以我在自我介绍的时候着重说了下自己的创业经历。

Link
Product — Q5 Estimate the distribution of time spent on market place. Y-axis is # of people, x axis is time spent on marketplace. (mean, median, mode)
应该是exponential distribution，大部分人只是浏览了5sec到10sec甚至更短，如何让浏览时间变长是产品的努力方向。




It's clear from the chart that the first 10 seconds of the page visit are critical for users' decision to stay or leave. The probability of leaving is very high during these first few seconds because

users are extremely skeptical, having suffered countless poorly designed web pages in the past. People know that most web pages are useless, and they behave accordingly to avoid wasting more time than absolutely necessary on bad pages.
If the web page survives this first — extremely harsh — 10-second judgment, users will look around a bit. However, they're still highly likely to leave during the subsequent 20 seconds of their visit. Only after people have stayed on a page for about 30 seconds does the curve become relatively flat. People continue to leave every second, but at a much slower rate than during the first 30 seconds.
So, if you can convince users to stay on your page for half a minute, there's a fair chance that they'll stay much longer — often 2 minutes or more, which is an eternity on the web.
So, roughly speaking, there are two cases here:

- bad pages, which get the chop in a few seconds; and
- good pages, which might be allocated a few minutes.

Note: "good" vs. "bad" is a decision that each individual user makes within those first few seconds of arriving. The design implications are clear:

- To gain several minutes of user attention, you must clearly communicate your value proposition within 10 seconds.

Q6. What is the benefit of launching Call to action (e.g., learn more, buy)? What metrics used for measure the effect? Want to sell?
Goal is to encourage people to sell products
Metrics:
# of users who sell their products
# Conversion rate (# people clicked yes/# people viewed)
# Monthly/daily active users
# Engagement/Time spent on browsing marketplace

Q7. The message_sent/session drops 10% month by month. What is the reason?
Replied Seasonality effect (look at historical data), time span(two months data vs. six months data), maybe buyer will not use facebook messenger to leave a note to buyers, competitors, does they show similar trend?
In the end, the interviewer mentioned it is mainly because area effects, some regions drops, while some regions increases. Need further investigation on this problem.

Link
产品：FB开发一个新产品，Pet page, 比如说Ins上有很多人专门给自己的宠物建一个账号，如果在FB上launch这个功能，让人们给自己的宠物建pet page， 请问怎么measure这个产品
产品我就回答了[td]adoption  perspective:  adoption rate, growth rate such as WoW and the trend plot
DAU， MAU，etc

## 10.Link Link2

SQL 是说friend request的2 tables
一个request table - (sender_id, send_to_id, time)
一个accept table - (requester_id, accepter_id, time)

Q1. On date XXX, what's the acceptance rate/percentage?

SELECT

```
        COUNT(A.requester_id)/COUNT(R.sender_id) AS acceptance_rate
FROM
        request R,
        accept A
WHERE
        DATE(R.time) = 'XXX' AND DATE(A.date) = 'XXX';
```

To remove duplicates:

```
SELECT
        COUNT(T2.requester_id)/COUNT(T1.sender_id) AS acceptance_rate
FROM
        (SELECT DISTINCT sender_id, send_to_id
         FROM request
         WHERE DATE(time) = 'XXX') T1,
        (SELECT DISTINCT requester_id, accepter_id
         FROM accept
         WHERE DATE(time) = 'XXX') T2;
```

Q2. 谁的friend 最多

```
SELECT
        user_id, SUM(num_friends) AS num_friends
FROM
        (SELECT requester_id AS user_id, COUNT(DISTINCT accepter_id) AS num_friends
         FROM accept
         GROUP BY 1
         UNION ALL
         SELECT accepter_id, COUNT(DISTINCT requester_id)
         FROM accept
         GROUP BY 1
        ) T1
GROUP BY 1
ORDER BY 2 DESC
LIMIT 1;
```

Product: FB打算有个feature, 给用户发text notification if their close friends update something. 这个feature有两步，一个问要不要sign up, 回答yes or no，说了yes就发。然后问如何决定要不要做这个feature，然后问了些metrics, 建议去看看AAREM这个framework，AAREM 是说一个 funnel process, Acquisition, Activation, Retention, Engagement, Monetization... 好像不同的人有不同的叫法，可以根据这个去查一下。

Link Link2

知道地里很多同学在准备product题目，其中实验设计经常出现，在这里想向大家请教两道题目，谢谢！

第一题：研究如果向用户推送close friends有update的notification，用户行为的变化

http://www.1point3acres.com/bbs/thread-273654-1-1.html

第二题：研究父母加入FB对用户行为的影响
http://www.1point3acres.com/bbs/thread-209706-1-1.html

## 我的问题是这两题应该用a/b testing还是cohort analysis?

a/b testing主要用于测试population里在同一时间一个变量对group A和group B的作用，group A和group B要comparable，要randomize。cohort analysis主要是分析population里面一小撮人，强调的是用户前后行为的对比。

a/b testing的问题：对于第一题，如果对愿意接受推送和不愿意的用户进行a/b testing是有bias的，愿意接收推送的人可能本来就很关心自己的close friends，所以有了推送之后，他们的活动会增加，这两个population就不是random的。对于第二题，如果直接比较有父母和没有父母的两组，很有可能是在有父母的里面，很多用户本来就不care父母，所以最后比较结果不显著。

Cohort的问题：没办法控制时间变量，可能有外在其他因素影响。

我的想法是，可否用两个cohort来进行前后对比。比如第一题，设置两个cohort，cohort A：接受的A，没接受的B。当cohort B在turn on前后没有显著差异的时候（确认没有外在条件影响），再分析cohort A在turn on前后是否有显著差异（但是感觉还是有点儿问题，比如如果外在事件只影响一个cohort呢）

Product — Close friend notification: 打算给用户的发text notification 告诉他们close friends 的Update, How to evaluate if we want to add this feature?
楼主回答说time spend on facebook, 看他没什么反应，然后又问这个notification是可以点击link到facebook上去的吗，他说是，于是我多嘴扯了一个CTR of the notification. 然后他追问你怎么evaluate CTR呢？然后我就马上跪了承认说这个metrics不好，没法draw the line，还是time spend on facebook吧.

Follow up: How to evaluate negative impact of this feature?
楼主问user是不是可以选择关闭这个服务，他说可以，楼主回答说看看percentage of user who close this feature

Follow up: If the engineer team roll out the feature on 1000 users, the time spend on facebook was 23 mins the week before roll out and 25 mins the week after. What would you say about this result?

楼主回答说先得rule out other factor may result the increase, do A/B test, 加个control to see if also see this increase trend.

Follow up: experiment 1000 user, 24 mins before, 26 mins after. Control arm 1000 user, 24 mins before, 24 mins after. What would you say about this result.
楼主脑子短路，在他提示variance是1.5mins，才反应过来这是在问hypothesis testing

不知道对不对啊求指正！

假设检验

uc: mean time spent on FB of the control group
ut: mean time spent on FB of the treatment group

构造了单边检验（这一点不太确定，感觉双边也可以？）

h0: uc = ut
ha: ut > uc

Assume alpha = 95%, 两组variance一样
用t检验
t =(ut-uc)/sqrt(var/1000+var/1000) = 2/sqrt(4.5/1000) = 30 >> 1.65
so reject h0

For the close friend notification question, I think we can break down the answer to three steps.

First, if the feature were successful, would it be good for Facebook? That is, would it move the key metric to the right direction? If we use engagement (time spent on Facebook, number of actions) as the metric, then the answer is yes. Sending text notification will potentially make people spend more time on Facebook and perform more actions (react to friends' posts).

Second, also the key to this question, is that we need to find a proxy for demand for this feature based on what users are doing today. The safest way for a data scientist to drive new features is to look at the data. If you find a demand for that feature (maybe users are going to friend's profile page to check for updates), then you can incentivize this behavior by simplifying the process. Sending text messages is a good way to do this.

Third, once you establish that the feature is good for Facebook's target metric and there's a demand for that today, we can run an A/B test to decide if we want to add it.

## **11.**Link Link2

SQL —
table 1: user1 | user2
123 456
456 123
123 789
789 123

table 2: sender | reciepient | action | date
123 456 create 2019-01-01
456 123 create 2019-01-01
123 789 create 2019-01-01

问每一对friends的interaction是多少？（一个create就是一个interaction）

```
SELECT t1.user1, t1.user2, count(t2.sender) as pair_total_interaction
FROM friend_pair t1
LEFT JOIN interaction t2
ON ((t1.user1 = t2.sender AND t1.user2 = t2.reciepient) OR (t1.user2 = t2.sender AND t1.user1
```

= t2.reciepient) )
group by 1,2

Product —
1. 一般说来一个user有很多friends是很好的，但是一个user有太多的friends也会有问题。你觉得会有什么问题?
(1) too many posts in the newsfeed, user may miss important updates from close friends who they care about most
(2) more spam or scams, higher chance that a user may add a friend which is a fake account sends them spams or scams

2. 如何来判断一个user的close friends
答：可以通过他们在一些特殊日子是不是会送礼物，照片里他们是不是会频繁出现被tag。
Best Friend: About Similarity, %Post/Photo Tagged Together, Messanger records

3. 如果有一个unfriend button， 如何向user推荐可以unfriend的人。
答： 可以给interaction（比如like，comment），互送gift， 照片里出现频率等factor加权重，然后算出一个score，根据score来排序。

Link
1.好友多是好事，但太多有时候也不好为什么
（好友太多，每天的post会太多，有些好友因为不熟并不关心他的动态，看不到想看的）
2. 基于你说的情况，我们称为crowded，那怎么定义这个存在crowded的situation。
（首先定义亲密好友，再比较亲密好友和非亲密好友的发帖比例）
3. 定义完亲密好友后，你打算用什么方法解决问题
推出unfriend的feature或者derank 非亲密好友的帖子
4.怎么定义亲密
共同好友，毕业学校，互动程度（互相点赞，互相发帖等等）然后weighted average出一个分数.
5.如果只有少部分人有互动怎么办
共同好友，text analtyics看post是否参加共同的活动，图像识别是否在同一张照片，是不是在同一个location
6. weighted average 的weight 怎么得到
我说可以用linear regression 或者其他machine learning mode得到
7.你要怎么implement这个model （当时楼主并没有意识到其实做supervised learning我们是没有y的，以为他问我怎么选模型，feauture engineering之类的）
楼主说了很多模型还有feature engineering以及evaluate模型的方法。。。
（被打断）8.用小白板指出没有y
我意识到问题，说那可以跟做survey得到y，或者用unsupervised model 做cluster
9.得到分数以后怎么做这个unfriend
我说根据分数升序排序推荐（分数越高越亲密）

## 12.Link

2 tabales: (1) date, userid, message_sends (2) date, userid, failed_message_sends. Write a query to obtain avg number of successful message-sends for users in 2 groups: (1)who did not have message_send failure on a given day. (2) users who faces message-send failure on that day)

(1)who did not have message_send failure on a given day

```
SELECT
        SUM(IFNULL(message_sends, 0))/COUNT(DISTINCT userid) AS avg_num
FROM
        table1
WHERE
        userid NOT IN (
        SELECT DISTINCT userid
        FROM table2
        )
        AND date = 'XXX';
```

(2) users who faces message-send failure on that day

```
SELECT
        IFNULL(SUM(T1.message_sends - IFNULL(T2.failed_message_sends, 0))/
COUNT(DISTINCT userid), 0) AS avg_num2
FROM
        table1 T1
        JOIN
        table2 T2
        ON T1.userid = T2.userid AND T1.date = T2.date
WHERE
        T1.date = 'XXX'
```

## 13.Link Link2

SQL spotify听歌问题，table1：timeluseridlsongidl table2: userid1luserid2

问(1)今天听的最频繁的歌曲是什么？(2)寻找一个list有userid和friendid：两个朋友有多于两首共同听过的歌曲

(1)

```
SELECT
        song_id, COUNT(user_id) AS num
FROM
        table1
WHERE DATE(time) = CURDATE()
GROUP BY 1
ORDER BY 2 DESC
LIMIT 1;
```

(2)

```
SELECT
        T1.userid1, T1.userid2, COUNT(DISTINCT T2.songid) AS num_song
FROM
        table2 T1
```

```
    JOIN table1 T2 ON T1.userid1 = T2.userid
    JOIN table1 T3 ON T1.userid2 = T3.userid AND T2.songid = T3.songid
GROUP BY 1, 2
HAVING num_song >= 2;
```

Product sense: 如果看到total ads revenue下降，你该怎么办？如何figure out？怎么给Senior management汇报
Ads revenue = # active users * CTR * price/click
Long term VS short term

## 14.Link

SQL部分 — Ad_ROI
Table1 adv_info:  advertiser_idl ad_idl spend: (The Advertiser pay for this ad)
Table2 ad_info:  ad_idl user_idl price: (The user spend through this ad (Assume all prices in this column >0))

Q1: The fraction of advertiser has at least 1 conversion?
Q2. What metrics would you show to advertisers (其实就是在问ROI), 用SQL实现

Q1:

```
SELECT
    COUNT(DISTINCT advertiser_id)/(SELECT COUNT(DISTINCT advertiser_id) FROM adv_info)
FROM
    adv_info
    JOIN
    ad_info
    ON ad_info.ad_id = adv_info.ad_id;
```

Q2: two cases
Case 1: advertiser每个ad的ROI

```
SELECT
    T1.advertiser_id, T1.ad_id, SUM(IFNULL(T2.price, 0))/T1.spend AS ad_ROI
FROM
    adv_info T1
    LEFT JOIN
    ad_info T2
    ON T1.ad_id = T2.ad_id
GROUP BY 1, 2;
```

Case 2: 每个advertiser的平均ROI

```
SELECT
    T1.advertiser_id, ROUND(IFNULL(total_revenue/total_spend, 0), 2) AS adv_ROI
FROM
    (SELECT advertiser_id, SUM(IFNULL(spend, 0)) AS total_spend
    FROM adv_info
    GROUP BY 1) T1
    LEFT JOIN
```

```
(SELECT advertiser_id, SUM(IFNULL(T2.price, 0)) AS total_revenue
FROM
        adv_info
        LEFT JOIN
        ad_info
        ON adv_info.ad_id = ad_info.ad_id
GROUP BY 1) T2
ON T1.advertiser_id = T2.advertiser_id;
```

## 15.Link Link2

SQL — survey_log 列名：user_id, question_id, question_order, event = {imp, answered, skipped}, timestamp,
第一问是找conversion rate最高的question，我在写完answer rate以后有个follow up问题，说如果imp太少的问题怎么办，我答的是加个threshold，舍弃那些出现少于多少次的问题（这里我随便写的10，他没有说什么，不知道对不对）。
第二问是在用户已经回答了某一问题的情况下，如何安排下一问题使conversion rate最高，我这里就按地里讨论的一样说在已经回答了这一问题的用户中，选他们回答的其余问题里回答率最高的一个

求了个每道题的回答率，注意不能只求回答次数，要除以总的看见此题的次数
追问，即使按照回答率对题目进行排序，如果新来的用户已经skip掉了回答率最高和次高的题，如何动态调整题目顺序，获得此用户尽可能多的回答?
我在这题上挣扎了很久，思考了题目内容分类，题目之间的相似度分类，等等等等，最后发现其实是条件概率，要看用户之间的相似度，即已有数据中跳过了这两道题的用户回答率最高的是哪道……

(1) 找conversion rate最高的question

```
SELECT
        question_id,
        SUM(CASE WHEN event = 'answered' THEN 1 ELSE 0 END)/SUM(CASE WHEN event = 'imp' THEN 1 ELSE 0 END) AS answer_rate
FROM
        survey_log
GROUP BY 1
ORDER BY 2 DESC
LIMIT 1;
```

(2)第二题，大概思路为找出在回答了现问题的人中，回答率最高的是那一个问题。因为，问题要知道两个变量，1，what's current question 2 what are questions the user  have answered? 下面subquery s找出了所有回答了current question 的人， subquery u找出了current user回答过的所有问题，因此，我们需要选出所有s 中回答率最高的问题，同时排除u中所有的问题。

```
#set the variables
set current_question = ?;
set current_user = ?;
#find the question_id that has the highest answer rate in the group that had answer current question
select question_id, sum(case when event = "answered" then 1 else 0 end)/sum(case when event = "imp"
```

```
then 1 else 0 end) as answer_rate
from
survey_log as L
join (select distinct user_id from survey_log where question_id = $current_question and event = "answer")
as s
on L.user_id = s.user_id
left join (select distinct question_id from survey_log where user_id = $current_user) as u
on L.question_id = u.question_id
where question_id != $current_question and u.question_id is null
order by answer_rate desc
limit 1;
```

Product

a. 先问how to measure health of facebook,我说可以分两个部分，一是website/sytem本身health，是不是function normally，另一部分是user engagement，比如新用户注册率，现有用户的interactions包括comment/share blablabla，然后DAU/MAU之类的，

b. 然后他直接打断说ok，现在就focus在comment上，如果要求你做一个dashboard specifically about comment，有哪些metrics你可以present。我先莫名想岔到之前面经的sql题了，跟他说可以show distribution of comments/users，然后他表现得不太满意，说ok，其他的呢，我就说可以present average comments per post/user，然后他继续问这个average comment per user是怎么计算的，denominator和numerator是啥，我回答numerator等于总评论数，denominator是distinct users who have left at least 1 comment per day。

c. 然后他说ok，现在把denominator换一换，改成DAU，我们发现这个总comment/DAU的比率跟一年前同一天相比，增加了50%,有哪些可能原因。我回答首先user mix可能有变化，可能今年新增的growth都是愿意发更多comment的用户，然后（随口乱说）content内容的变化吸引更多user leave comment，第三可能有新feature刺激用户留言更多之类的。然后他follow up了一句什么我忘了，好像是如何判断是哪种？我就先说我们可以segmentation by users' language/platform/device type/browser/location看看有没有哪个subgroups特别突出，有的话可以dive deep。

SQL Questions

# Questions

## 1. Posts

Table: user_actions
date | actor_id | post_id | relationship | interaction

Table: user_posts
date | poster_id | post_id

(1) How many likes were made on Friend posts yesterday?

(2) If I were user 123, how would you calculate the average number of likes on all of my posts?

## 2. Friending

Table: friends
sender_id | receiver_id | sent_date | accepted_date | sender_country

Table: spams
userid | spam_type

Table: age_group
userid | age_group

(3) What is the same day acceptance rate for every day in the last week?

(4) Find pairs of users who had no interaction last year.

(5) What is the accepted rate per day from the sender last week?

(6) What is the percent of fraud users that sent requests in the last 7 days?
(7) What are the total friend requests for each day of week in the past 4 weeks?
(8) What is the average sent request for each age group in the past week?

## 3. Search

Table: searches
date | search_id ( the unique identifier of each) | user_id | age_group ('U30', '30-50', '50+') | search_query
‘2020-01-01’ | 101 | 9991 | 'U30' | ‘michael jackson’
Table: search_results
date | search_id | result_id | result_type | clicked
‘2020-01-01’ | 101 | 1001 | ‘page’ | TRUE
‘2020-01-01’ | 101 | 1002 | ‘event’ | FALSE
‘2020-01-01’ | 101 | 1004 | ‘group' | FALSE

(9) How many users searched for "john" in the last 7 days?

(10) % user click events after search?

(11) by each age group, how many unique users search for “dog” in the last 7 days??

(12) What are the top 10 search terms that are most likely to return at least one result about an event?
(13) In the last 7 days, how many users conducted more than 10 searches?
(14) How many users conducted searches that lead to multiple result_type?
(15) 过去一周，search_query top 10
(16) how many users searched event in the past 30 days
(17) What% of users clicked on a result about an Event?
(18) How many users searched "dog" more than 10 times in the last 7 days?
(19) Which result type has most clicks within each age group for the search query "dog"?

## 4. Video

Table: video_calls
caller| recipient| date| call_id| duration

Table: fb_dau
user_id| DAU_flag| date| country

(20) On '2020-01-01' how many people initiate multiple calls?

(21) % of DAU used the video calls functions on '2020-01-01' in France?
(22) 这个月内排名前三initialize 最多通电话的用户?
(23) % of DAU used the video calls functions by each country on '2020-01-01'?

## 5. Spam / Reported Posts

Table: user_actions
ds (STRING) | user_id (BIGINT) |post_id (BIGINT) |action (STRING) | extra (STRING)
'2018-07-01'| 1209283021 | 329482048384792 | 'view' |
'2018-07-01'| 1209283021 | 329482048384792 | 'like' |
'2018-07-01'| 1938409273 | 349573908750923 | 'reaction' | 'LOVE'
'2018-07-01'| 1209283021 | 329482048384792 | 'comment' | 'Such nice Raybans'
'2018-07-01'| 1238472931 | 329482048384792 | 'report' | 'SPAM'

Table: reviewer_removals
ds (STRING) | reviewer_id (BIGINT) | post_id (BIGINT)
'2018-07-01'| 3894729384729078 | 329482048384792 |

(24) How many posts were reported yesterday for each report reason?

(25) What percent of daily content that users view on Facebook is actually Spam?
(26) 昨天有多少条post被report spam了？
(27) 算一下每一条post 被举报为spam的百分比？ 也就是每条post的#report spam／#view

## 6. Friend Requests

Table: user_actions
date | actor_id | post_id | relationship | interaction

Table: user_posts
date | poster_id | post_id

(28) How many likes were made on Friend posts yesterday?

(29) If I were user 123, how would you calculate the average number of likes on all of my posts?

## 7. Ads

Table: Ads
ad_id | user_id | session_id | action ('viewed', 'clicked', 'hide') | date

(30) What do you think is the best performing ad in the past 10 days? (discuss definition of best performing, # clicks or ctr)
(31) From this info, how to recommend the next ad to the user given they viewed a particular ad?
(32) Average time after people view a ad to click.

## 8. Marketplace

Table: commerce_user_actions
sessionid | userid | date | event (surface_enter, surface_exit, click, first_scroll, message_send)
*Only surface_enter happens for every session. Other events may or may not happen. For example, if user sessions times out, there won't be a surface exit

Table: time_spent_per_session
date | sessionid | time_spent_sec

(33) Calculate the average number of sessions/user by day for the last 30 days
(34) Think about the time spent distribution on the marketplace by users for a day or on '2018-08-01'. (x-axis is its bucket and y-axis is the number of users). ***What may the distribution look like?*.

# Useful SQL questions from Leetcode

**176. Second Highest Salary**
**Write a SQL query to get the second highest salary from the Employee table.**

```
+----+--------+
| Id | Salary |
+----+--------+
| 1  | 100    |
```

| 2  | 200    |
| 3  | 300    |
+----+--------+
For example, given the above Employee table, the query should return 200 as the second highest salary. If there is no second highest salary, then the query should return null.

**178. Rank Scores**
Write a SQL query to rank scores. If there is a tie between two scores, both should have the same ranking. Note that after a tie, the next ranking number should be the next consecutive integer value. In other words, there should be no "holes" between ranks.
+----+-------+
| Id | Score |
+----+-------+
| 1  | 3.50  |
| 2  | 3.65  |
| 3  | 4.00  |
| 4  | 3.85  |
| 5  | 4.00  |
| 6  | 3.65  |
+----+-------+
For example, given the above Scores table, your query should generate the following report (order by highest score):
+-------+------+
| Score | Rank |
+-------+------+
| 4.00  | 1    |
| 4.00  | 1    |
| 3.85  | 2    |
| 3.65  | 3    |
| 3.65  | 3    |
| 3.50  | 4    |
+-------+------+

**180. Consecutive Numbers**
Write a SQL query to find all numbers that appear at least three times consecutively.
+----+-----+
| Id | Num |
+----+-----+
| 1  |  1  |
| 2  |  1  |
| 3  |  1  |
| 4  |  2  |
| 5  |  1  |
| 6  |  2  |
| 7  |  2  |
+----+-----+
For example, given the above Logs table, 1 is the only number that appears consecutively for at least three times.
+-----------------+
| ConsecutiveNums |
+-----------------+

| 1 |

**182. Duplicate Emails**

Write a SQL query to find all duplicate emails in a table named Person.

```
+----+---------+
| Id | Email   |
+----+---------+
| 1  | a@b.com |
| 2  | c@d.com |
| 3  | a@b.com |
+----+---------+
```

For example, your query should return the following for the above table:

```
+---------+
| Email   |
+---------+
| a@b.com |
+---------+
```

**183. Customers Who Never Order**

Suppose that a website contains two tables, the Customers table and the Orders table. Write a SQL query to find all customers who never order anything.

Table: Customers.

```
+----+-------+
| Id | Name  |
+----+-------+
| 1  | Joe   |
| 2  | Henry |
| 3  | Sam   |
| 4  | Max   |
+----+-------+
```

Table: Orders.

```
+----+------------+
| Id | CustomerId |
+----+------------+
| 1  | 3          |
| 2  | 1          |
+----+------------+
```

Using the above tables as example, return the following:

```
+-----------+
| Customers |
+-----------+
| Henry     |
| Max       |
+-----------+
```

**534. Game Play Analysis III (SUM OVER WINDOW)**

**Table: Activity**

+--------------+---------+
| Column Name  | Type    |
+--------------+---------+
| player_id    | int     |
| device_id    | int     |
| event_date   | date    |
| games_played | int     |
+--------------+---------+
(player_id, event_date) is the primary key of this table.
This table shows the activity of players of some game.
Each row is a record of a player who logged in and played a number of games (possibly 0) before logging out on some day using some device.

**Write an SQL query that reports for each player and date, how many games played so far by the player. That is, the total number of games played by the player until that date. Check the example for clarity.**

**The query result format is in the following example:**

Activity table:
+-----------+-----------+------------+--------------+
| player_id | device_id | event_date | games_played |
+-----------+-----------+------------+--------------+
| 1         | 2         | 2016-03-01 | 5            |
| 1         | 2         | 2016-05-02 | 6            |
| 1         | 3         | 2017-06-25 | 1            |
| 3         | 1         | 2016-03-02 | 0            |
| 3         | 4         | 2018-07-03 | 5            |
+-----------+-----------+------------+--------------+

Result table:
+-----------+------------+---------------------+
| player_id | event_date | games_played_so_far |
+-----------+------------+---------------------+
| 1         | 2016-03-01 | 5                   |
| 1         | 2016-05-02 | 11                  |
| 1         | 2017-06-25 | 12                  |
| 3         | 2016-03-02 | 0                   |
| 3         | 2018-07-03 | 5                   |
+-----------+------------+---------------------+
For the player with id 1, 5 + 6 = 11 games played by 2016-05-02, and 5 + 6 + 1 = 12 games played by 2017-06-25.
For the player with id 3, 0 + 5 = 5 games played by 2018-07-03.
Note that for each player we only care about the days when the player logged in.

**550. Game Play Analysis IV**

Table: Activity
+--------------+---------+
| Column Name  | Type    |
+--------------+---------+
| player_id    | int     |

| device_id | int |
| event_date | date |
| games_played | int |
+--------------+---------+
(player_id, event_date) is the primary key of this table.
This table shows the activity of players of some game.
Each row is a record of a player who logged in and played a number of games (possibly 0) before logging out on some day using some device.

Write an SQL query that reports the **fraction** of players that logged in again on the day after the day they first logged in, **rounded to 2 decimal places**. In other words, you need to count the number of players that logged in for at least two consecutive days starting from their first login date, then divide that number by the total number of players.
The query result format is in the following example:
Activity table:
+-----------+-----------+------------+--------------+
| player_id | device_id | event_date | games_played |
+-----------+-----------+------------+--------------+
| 1 | 2 | 2016-03-01 | 5 |
| 1 | 2 | 2016-03-02 | 6 |
| 2 | 3 | 2017-06-25 | 1 |
| 3 | 1 | 2016-03-02 | 0 |
| 3 | 4 | 2018-07-03 | 5 |
+-----------+-----------+------------+--------------+

Result table:
+-----------+
| fraction |
+-----------+
| 0.33 |
+-----------+
Only the player with id 1 logged back in after the first day he had logged in so the answer is 1/3 = 0.33

**569. Median Employee Salary**
The Employee table holds all employees. The employee table has three columns: Employee Id, Company Name, and Salary.
+-----+------------+--------+
|Id | Company | Salary |
+-----+------------+--------+
|1 | A | 2341 |
|2 | A | 341 |
|3 | A | 15 |
|4 | A | 15314 |
|5 | A | 451 |
|6 | A | 513 |
|7 | B | 15 |
|8 | B | 13 |
|9 | B | 1154 |
|10 | B | 1345 |
|11 | B | 1221 |

|12 | B | 234 |
|13 | C | 2345 |
|14 | C | 2645 |
|15 | C | 2645 |
|16 | C | 2652 |
|17 | C | 65 |
+-----+------------+--------+

Write a SQL query to find the median salary of each company. Bonus points if you can solve it without using any built-in SQL functions.

+-----+------------+--------+
|Id | Company | Salary |
+-----+------------+--------+
|5 | A | 451 |
|6 | A | 513 |
|12 | B | 234 |
|9 | B | 1154 |
|14 | C | 2645 |
+-----+------------+--------+

**574. Winning Candidate**

## Table: Candidate

+-----+---------+
| id | Name |
+-----+---------+
| 1 | A |
| 2 | B |
| 3 | C |
| 4 | D |
| 5 | E |
+-----+---------+

## Table: Vote

+-----+--------------+
| id | CandidateId |
+-----+--------------+
| 1 | 2 |
| 2 | 4 |
| 3 | 3 |
| 4 | 2 |
| 5 | 5 |
+-----+--------------+

id is the auto-increment primary key,
CandidateId is the id appeared in Candidate table.

## Write a sql to find the name of the winning candidate, the above example will return the winner B.

+------+
| Name |
+------+
| B |

+------+

## Notes:

1. You may assume **there is no tie**, in other words there will be **only one** winning candidate.
2.

**579. Find Cumulative Salary of an Employee**

The **Employee** table holds the salary information in a year.

Write a SQL to get the cumulative sum of an employee's salary over a period of 3 months but exclude the most recent month.

The result should be displayed by 'Id' ascending, and then by 'Month' descending.

**Example**

**Input**

| Id | Month | Salary |
|---|---|---|
| 1 | 1 | 20 |
| 2 | 1 | 20 |
| 1 | 2 | 30 |
| 2 | 2 | 30 |
| 3 | 2 | 40 |
| 1 | 3 | 40 |
| 3 | 3 | 60 |
| 1 | 4 | 60 |
| 3 | 4 | 70 |

**Output**

| Id | Month | Salary |
|---|---|---|
| 1 | 3 | 90 |
| 1 | 2 | 50 |
| 1 | 1 | 20 |
| 2 | 1 | 20 |
| 3 | 3 | 100 |
| 3 | 2 | 40 |

**Explanation**

Employee '1' has 3 salary records for the following 3 months except the most recent month '4': salary 40 for month '3', 30 for month '2' and 20 for month '1'

So the cumulative sum of salary of this employee over 3 months is 90(40+30+20), 50(30+20) and 20 respectively.

| Id | Month | Salary |
|---|---|---|
| 1 | 3 | 90 |
| 1 | 2 | 50 |
| 1 | 1 | 20 |

Employee '2' only has one salary record (month '1') except its most recent month '2'.

| Id | Month | Salary |
|---|---|---|
| 2 | 1 | 20 |

**1098. Unpopular Books**

Table: Books

```
+----------------+---------+
| Column Name    | Type    |
+----------------+---------+
| book_id        | int     |
| name           | varchar |
| available_from | date    |
+----------------+---------+
```

book_id is the primary key of this table.

Table: Orders

```
+----------------+---------+
| Column Name    | Type    |
+----------------+---------+
| order_id       | int     |
| book_id        | int     |
| quantity       | int     |
| dispatch_date  | date    |
+----------------+---------+
```

order_id is the primary key of this table.
book_id is a foreign key to the Books table.

Write an SQL query that reports the **books** that have sold **less than 10** copies in the last year, excluding books that have been available for less than 1 month from today. **Assume today is 2019-06-23**.

The query result format is in the following example:

Books table:

```
+---------+--------------------+----------------+
| book_id | name               | available_from |
+---------+--------------------+----------------+
| 1       | "Kalila And Demna" | 2010-01-01     |
| 2       | "28 Letters"       | 2012-05-12     |
| 3       | "The Hobbit"       | 2019-06-10     |
| 4       | "13 Reasons Why"   | 2019-06-01     |
| 5       | "The Hunger Games" | 2008-09-21     |
+---------+--------------------+----------------+
```

Orders table:

```
+----------+---------+----------+---------------+
| order_id | book_id | quantity | dispatch_date |
+----------+---------+----------+---------------+
| 1        | 1       | 2        | 2018-07-26    |
| 2        | 1       | 1        | 2018-11-05    |
| 3        | 3       | 8        | 2019-06-11    |
| 4        | 4       | 6        | 2019-06-05    |
| 5        | 4       | 5        | 2019-06-20    |
| 6        | 5       | 9        | 2009-02-02    |
| 7        | 5       | 8        | 2010-04-13    |
+----------+---------+----------+---------------+
```

Result table:

+----------+--------------------+
| book_id  | name               |
+----------+--------------------+
| 1        | "Kalila And Demna" |
| 2        | "28 Letters"       |
| 5        | "The Hunger Games" |
+----------+--------------------+

**1112. Highest Grade For Each Student**

Table: Enrollments

+---------------+---------+
| Column Name   | Type    |
+---------------+---------+
| student_id    | int     |
| course_id     | int     |
| grade         | int     |
+---------------+---------+

(student_id, course_id) is the primary key of this table.

Write a SQL query to find the highest grade with its corresponding course for each student. In case of a tie, you should find the course with the smallest course_id. The output must be sorted by increasing student_id.

The query result format is in the following example:

Enrollments table:

+------------+-----------+-------+
| student_id | course_id | grade |
+------------+-----------+-------+
| 2          | 2         | 95    |
| 2          | 3         | 95    |
| 1          | 1         | 90    |
| 1          | 2         | 99    |
| 3          | 1         | 80    |
| 3          | 2         | 75    |
| 3          | 3         | 82    |
+------------+-----------+-------+

Result table:

+------------+-----------+-------+
| student_id | course_id | grade |
+------------+-----------+-------+
| 1          | 2         | 99    |
| 2          | 2         | 95    |
| 3          | 3         | 82    |
+------------+-----------+-------+

**1141. User Activity for the Past 30 Days I**

**Table: Activity**

+---------------+---------+
| Column Name   | Type    |
+---------------+---------+

| user_id | int |
| session_id | int |
| activity_date | date |
| activity_type | enum |
+---------------+---------+
There is no primary key for this table, it may have duplicate rows.
The activity_type column is an ENUM of type ('open_session', 'end_session', 'scroll_down', 'send_message').
The table shows the user activities for a social media website.
Note that each session belongs to exactly one user.

**Write an SQL query to find the daily active user count for a period of 30 days ending 2019-07-27 inclusively. A user was active on some day if he/she made at least one activity on that day.**

**The query result format is in the following example:**

Activity table:
+---------+------------+---------------+---------------+
| user_id | session_id | activity_date | activity_type |
+---------+------------+---------------+---------------+
| 1 | 1 | 2019-07-20 | open_session |
| 1 | 1 | 2019-07-20 | scroll_down |
| 1 | 1 | 2019-07-20 | end_session |
| 2 | 4 | 2019-07-20 | open_session |
| 2 | 4 | 2019-07-21 | send_message |
| 2 | 4 | 2019-07-21 | end_session |
| 3 | 2 | 2019-07-21 | open_session |
| 3 | 2 | 2019-07-21 | send_message |
| 3 | 2 | 2019-07-21 | end_session |
| 4 | 3 | 2019-06-25 | open_session |
| 4 | 3 | 2019-06-25 | end_session |
+---------+------------+---------------+---------------+

Result table:
+------------+--------------+
| day | active_users |
+------------+--------------+
| 2019-07-20 | 2 |
| 2019-07-21 | 2 |
+------------+--------------+
Note that we do not care about days with zero active users.

**1142. User Activity for the Past 30 Days II**

**Table: Activity**

+---------------+---------+
| Column Name | Type |
+---------------+---------+
| user_id | int |
| session_id | int |
| activity_date | date |
| activity_type | enum |

+---------------+---------+
There is no primary key for this table, it may have duplicate rows.
The activity_type column is an ENUM of type ('open_session', 'end_session', 'scroll_down', 'send_message').
The table shows the user activities for a social media website.
Note that each session belongs to exactly one user.

**Write an SQL query to find the average number of sessions per user for a period of 30 days ending 2019-07-27 inclusively, rounded to 2 decimal places. The sessions we want to count for a user are those with at least one activity in that time period.**

**The query result format is in the following example:**

Activity table:

| user_id | session_id | activity_date | activity_type |
|---|---|---|---|
| 1 | 1 | 2019-07-20 | open_session |
| 1 | 1 | 2019-07-20 | scroll_down |
| 1 | 1 | 2019-07-20 | end_session |
| 2 | 4 | 2019-07-20 | open_session |
| 2 | 4 | 2019-07-21 | send_message |
| 2 | 4 | 2019-07-21 | end_session |
| 3 | 2 | 2019-07-21 | open_session |
| 3 | 2 | 2019-07-21 | send_message |
| 3 | 2 | 2019-07-21 | end_session |
| 3 | 5 | 2019-07-21 | open_session |
| 3 | 5 | 2019-07-21 | scroll_down |
| 3 | 5 | 2019-07-21 | end_session |
| 4 | 3 | 2019-06-25 | open_session |
| 4 | 3 | 2019-06-25 | end_session |

Result table:

| average_sessions_per_user |
|---|
| 1.33 |

User 1 and 2 each had 1 session in the past 30 days while user 3 had 2 sessions so the average is (1 + 1 + 2) / 3 = 1.33.

## 1158. Market Analysis I

### Table: Users

| Column Name | Type |
|---|---|
| user_id | int |
| join_date | date |
| favorite_brand | varchar |

user_id is the primary key of this table.

This table has the info of the users of an online shopping website where users can sell and buy items.

### Table: Orders

```
+---------------+---------+
| Column Name   | Type    |
+---------------+---------+
| order_id      | int     |
| order_date    | date    |
| item_id       | int     |
| buyer_id      | int     |
| seller_id     | int     |
+---------------+---------+
```

order_id is the primary key of this table.
item_id is a foreign key to the Items table.
buyer_id and seller_id are foreign keys to the Users table.

### Table: Items

```
+---------------+---------+
| Column Name   | Type    |
+---------------+---------+
| item_id       | int     |
| item_brand    | varchar |
+---------------+---------+
```

item_id is the primary key of this table.

### Write an SQL query to find for each user, the join date and the number of orders they made as a buyer in 2019.

### The query result format is in the following example:

Users table:

```
+---------+------------+----------------+
| user_id | join_date  | favorite_brand |
+---------+------------+----------------+
| 1       | 2018-01-01 | Lenovo         |
| 2       | 2018-02-09 | Samsung        |
| 3       | 2018-01-19 | LG             |
| 4       | 2018-05-21 | HP             |
+---------+------------+----------------+
```

Orders table:

```
+----------+------------+---------+----------+-----------+
| order_id | order_date | item_id | buyer_id | seller_id |
+----------+------------+---------+----------+-----------+
| 1        | 2019-08-01 | 4       | 1        | 2         |
| 2        | 2018-08-02 | 2       | 1        | 3         |
| 3        | 2019-08-03 | 3       | 2        | 3         |
| 4        | 2018-08-04 | 1       | 4        | 2         |
| 5        | 2018-08-04 | 1       | 3        | 4         |
| 6        | 2019-08-05 | 2       | 2        | 4         |
+----------+------------+---------+----------+-----------+
```

Items table:

```
+---------+------------+
| item_id | item_brand |
+---------+------------+
| 1       | Samsung    |
| 2       | Lenovo     |
| 3       | LG         |
| 4       | HP         |
+---------+------------+
```

**1211. Queries Quality and Percentage**

**Table: Queries**

```
+-------------+---------+
| Column Name | Type    |
+-------------+---------+
| query_name  | varchar |
| result      | varchar |
| position    | int     |
| rating      | int     |
+-------------+---------+
```

There is no primary key for this table, it may have duplicate rows.
This table contains information collected from some queries on a database.
The position column has a value from **1** to **500**.
The rating column has a value from **1** to **5**. Query with rating less than 3 is a poor query.

We define query quality as:
The average of the ratio between query rating and its position.
We also define poor query percentage as:
The percentage of all queries with rating less than 3.
Write an SQL query to find each query_name, the quality and poor_query_percentage.
Both quality and poor_query_percentage should be rounded to 2 decimal places.
The query result format is in the following example:

Queries table:

```
+------------+-------------------+----------+--------+
| query_name | result            | position | rating |
+------------+-------------------+----------+--------+
| Dog        | Golden Retriever  | 1        | 5      |
| Dog        | German Shepherd   | 2        | 5      |
| Dog        | Mule              | 200      | 1      |
| Cat        | Shirazi           | 5        | 2      |
| Cat        | Siamese           | 3        | 3      |
| Cat        | Sphynx            | 7        | 4      |
+------------+-------------------+----------+--------+
```

Result table:

```
+------------+---------+-----------------------+
| query_name | quality | poor_query_percentage |
+------------+---------+-----------------------+
| Dog        | 2.50    | 33.33                 |
```

| Cat | 0.66 | 33.33 |
+------------+---------+----------------------+

Dog queries quality is ((5 / 1) + (5 / 2) + (1 / 200)) / 3 = 2.50
Dog queries poor_ query_percentage is (1 / 3) * 100 = 33.33

Cat queries quality equals ((2 / 5) + (3 / 3) + (4 / 7)) / 3 = 0.66
Cat queries poor_ query_percentage is (1 / 3) * 100 = 33.33

**1241. Number of Comments per Post**

## Table: Submissions

+---------------+----------+
| Column Name | Type |
+---------------+----------+
| sub_id | int |
| parent_id | int |
+---------------+----------+
There is no primary key for this table, it may have duplicate rows.
Each row can be a post or comment on the post.
parent_id is null for posts.
parent_id for comments is sub_id for another post in the table.

**Write an SQL query to find number of comments per each post.**

**Result table should contain post_id and its corresponding number_of_comments, and must be sorted by post_id in ascending order.**

**Submissions may contain duplicate comments. You should count the number of unique comments per post.**

**Submissions may contain duplicate posts. You should treat them as one post.**

**The query result format is in the following example:**

Submissions table:
+---------+------------+
| sub_id | parent_id |
+---------+------------+
| 1 | Null |
| 2 | Null |
| 1 | Null |
| 12 | Null |
| 3 | 1 |
| 5 | 2 |
| 3 | 1 |
| 4 | 1 |
| 9 | 1 |
| 10 | 2 |
| 6 | 7 |
+---------+------------+

Result table:
+---------+--------------------+

| post_id | number_of_comments |
+---------+--------------------+
| 1       | 3                  |
| 2       | 2                  |
| 12      | 0                  |
+---------+--------------------+

The post with id 1 has three comments in the table with id 3, 4 and 9. The comment with id 3 is repeated in the table, we counted it **only once**.
The post with id 2 has two comments in the table with id 5 and 10.
The post with id 12 has no comments in the table.
The comment with id 6 is a comment on a deleted post with id 7 so we ignored it.

**1264. Page Recommendations**

### Table: Friendship

+---------------+---------+
| Column Name   | Type    |
+---------------+---------+
| user1_id      | int     |
| user2_id      | int     |
+---------------+---------+
(user1_id, user2_id) is the primary key for this table.
Each row of this table indicates that there is a friendship relation between user1_id and user2_id.

### Table: Likes

+-------------+---------+
| Column Name | Type    |
+-------------+---------+
| user_id     | int     |
| page_id     | int     |
+-------------+---------+
(user_id, page_id) is the primary key for this table.
Each row of this table indicates that user_id likes page_id.

### Write an SQL query to recommend pages to the user with user_id = 1 using the pages that your friends liked. It should not recommend pages you already liked.

### Return result table in any order without duplicates.

### The query result format is in the following example:

Friendship table:
+----------+----------+
| user1_id | user2_id |
+----------+----------+
| 1        | 2        |
| 1        | 3        |
| 1        | 4        |
| 2        | 3        |
| 2        | 4        |

| 2 | 5 |
| 6 | 1 |
+----------+----------+

Likes table:
+---------+---------+
| user_id | page_id |
+---------+---------+
| 1 | 88 |
| 2 | 23 |
| 3 | 24 |
| 4 | 56 |
| 5 | 11 |
| 6 | 33 |
| 2 | 77 |
| 3 | 77 |
| 6 | 88 |
+---------+---------+

Result table:
+------------------+
| recommended_page |
+------------------+
| 23 |
| 24 |
| 56 |
| 33 |
| 77 |
+------------------+
User one is friend with users 2, 3, 4 and 6.
Suggested pages are 23 from user 2, 24 from user 3, 56 from user 3 and 33 from user 6.
Page 77 is suggested from both user 2 and user 3.
Page 88 is not suggested because user 1 already likes it.

**1322. Ads Performance**

## Table: Ads

+---------------+---------+
| Column Name | Type |
+---------------+---------+
| ad_id | int |
| user_id | int |
| action | enum |
+---------------+---------+
(ad_id, user_id) is the primary key for this table.
Each row of this table contains the ID of an Ad, the ID of a user and the action taken by this user regarding this Ad.
The action column is an ENUM type of ('Clicked', 'Viewed', 'Ignored').

### A company is running Ads and wants to calculate the performance of each Ad.

**Performance of the Ad is measured using Click-Through Rate (CTR) where:**

**Write an SQL query to find the ctr of each Ad.**

**Round ctr to 2 decimal points. Order the result table by ctr in descending order and by ad_id in ascending order in case of a tie.**

**The query result format is in the following example:**

Ads table:

```
+-------+---------+---------+
| ad_id | user_id | action  |
+-------+---------+---------+
| 1     | 1       | Clicked |
| 2     | 2       | Clicked |
| 3     | 3       | Viewed  |
| 5     | 5       | Ignored |
| 1     | 7       | Ignored |
| 2     | 7       | Viewed  |
| 3     | 5       | Clicked |
| 1     | 4       | Viewed  |
| 2     | 11      | Viewed  |
| 1     | 2       | Clicked |
+-------+---------+---------+
```

Result table:

```
+-------+-------+
| ad_id | ctr   |
+-------+-------+
| 1     | 66.67 |
| 3     | 50.00 |
| 2     | 33.33 |
| 5     | 0.00  |
+-------+-------+
```

for ad_id = 1, ctr = (2/(2+1)) * 100 = 66.67
for ad_id = 2, ctr = (1/(1+2)) * 100 = 33.33
for ad_id = 3, ctr = (1/(1+1)) * 100 = 50.00
for ad_id = 5, ctr = 0.00, Note that ad_id = 5 has no clicks or views.
Note that we don't care about Ignored Ads.
Result table is ordered by the ctr. in case of a tie we order them by ad_id

# Tips

## Change numeric type to character

```
SELECT CAST(funding_total_usd AS varchar) AS funding_total_usd_string,
founded_at_clean::varchar AS founded_at_string FROM
tutorial.crunchbase_companies_clean_date
```

Date Format

```
SELECT companies.category_code, COUNT(CASE WHEN acquisitions.acquired_at_cleaned <=
companies.founded_at_clean::timestamp + INTERVAL '3 years' THEN 1 ELSE NULL END) AS
acquired_3_yrs, COUNT(CASE WHEN acquisitions.acquired_at_cleaned <=
companies.founded_at_clean::timestamp + INTERVAL '5 years' THEN 1 ELSE NULL END) AS
acquired_5_yrs, COUNT(CASE WHEN acquisitions.acquired_at_cleaned <=
companies.founded_at_clean::timestamp + INTERVAL '10 years' THEN 1 ELSE NULL END) AS
acquired_10_yrs, COUNT(1) AS total FROM tutorial.crunchbase_companies_clean_date
companies JOIN tutorial.crunchbase_acquisitions_clean_date acquisitions ON
acquisitions.company_permalink = companies.permalink WHERE founded_at_clean IS NOT
NULL GROUP BY 1 ORDER BY 5 DESC
```

String Format

```
SELECT incidnt_num,
       date,
       LEFT(date, 10) AS cleaned_date,
       RIGHT(date, LENGTH(date) - 11) AS cleaned_time
  FROM tutorial.sf_crime_incidents_2014_01

SELECT location,
       TRIM(both '()' FROM location)
  FROM tutorial.sf_crime_incidents_2014_01
```

The `TRIM` function takes 3 arguments. First, you have to specify whether you want to remove characters from the beginning ('leading'), the end ('trailing'), or both ('both', as used above). Next you must specify all characters to be trimmed. Any characters included in the single quotes will be removed from both beginning, end, or both sides of the string. Finally, you must specify the text you want to trim using `FROM`.

```
SELECT incidnt_num,
       descript,
       POSITION('A' IN descript) AS a_position
  FROM tutorial.sf_crime_incidents_2014_01

SELECT incidnt_num,
       date,
       SUBSTR(date, 4, 2) AS day
  FROM tutorial.sf_crime_incidents_2014_01

SELECT incidnt_num,
       day_of_week,
       LEFT(date, 10) AS cleaned_date,
       CONCAT(day_of_week, ', ', LEFT(date, 10)) AS day_and_date
```

```
  FROM tutorial.sf_crime_incidents_2014_01
```

```
SELECT location,
TRIM(leading '(' FROM LEFT(location, POSITION(',' IN location) - 1)) AS lattitude,
TRIM(trailing ')' FROM RIGHT(location, LENGTH(location) - POSITION(',' IN location) )
) AS longitude
FROM tutorial.sf_crime_incidents_2014_01
```

with the first letter capitalized and the rest of the letters in lower-case

```
SELECT incidnt_num, category,
UPPER(LEFT(category, 1)) || LOWER(RIGHT(category, LENGTH(category) - 1)) AS
category_cleaned
FROM tutorial.sf_crime_incidents_2014_01
```

```
SELECT cleaned_date,
       EXTRACT('year'    FROM cleaned_date) AS year,
       EXTRACT('month'   FROM cleaned_date) AS month,
       EXTRACT('day'     FROM cleaned_date) AS day,
       EXTRACT('hour'    FROM cleaned_date) AS hour,
       EXTRACT('minute'  FROM cleaned_date) AS minute,
       EXTRACT('second'  FROM cleaned_date) AS second,
       EXTRACT('decade'  FROM cleaned_date) AS decade,
       EXTRACT('dow'     FROM cleaned_date) AS day_of_week
  FROM tutorial.sf_crime_incidents_cleandate
```

This happens frequently in numerical data (displaying nulls as 0 is often preferable), and when performing outer joins that result in some unmatched rows. In cases like this, you can use `COALESCE` to replace the null values:

```
SELECT incidnt_num,
       descript,
       COALESCE(descript, 'No Description')
  FROM tutorial.sf_crime_incidents_cleandate
 ORDER BY descript DESC
```

Window functions: https://mode.com/sql-tutorial/sql-window-functions/

The most practical example of this is a running total:

```
SELECT duration_seconds,
       SUM(duration_seconds) OVER (ORDER BY start_time) AS running_total
  FROM tutorial.dc_bikeshare_q1_2012
```

```
SELECT start_terminal,
       duration_seconds,
       SUM(duration_seconds) OVER
         (PARTITION BY start_terminal) AS running_total,
       COUNT(duration_seconds) OVER
         (PARTITION BY start_terminal) AS running_count,
       AVG(duration_seconds) OVER
         (PARTITION BY start_terminal) AS running_avg
  FROM tutorial.dc_bikeshare_q1_2012
 WHERE start_time < '2012-01-08'
```

| | start_terminal | duration_seconds | running_total | running_count | running_avg |
|---|---|---|---|---|---|
| 1 | 31000 | 277 | 12207 | 16 | 762.9375 |
| 2 | 31000 | 1422 | 12207 | 16 | 762.9375 |
| 3 | 31000 | 398 | 12207 | 16 | 762.9375 |
| 4 | 31000 | 414 | 12207 | 16 | 762.9375 |
| 5 | 31000 | 3340 | 12207 | 16 | 762.9375 |
| 6 | 31000 | 291 | 12207 | 16 | 762.9375 |
| 7 | 31000 | 2661 | 12207 | 16 | 762.9375 |

```
SELECT start_terminal,
       duration_seconds,
       SUM(duration_seconds) OVER
         (PARTITION BY start_terminal ORDER BY start_time)
         AS running_total,
       COUNT(duration_seconds) OVER
         (PARTITION BY start_terminal ORDER BY start_time)
         AS running_count,
       AVG(duration_seconds) OVER
         (PARTITION BY start_terminal ORDER BY start_time)
         AS running_avg
  FROM tutorial.dc_bikeshare_q1_2012
 WHERE start_time < '2012-01-08'
```

| | start_terminal | duration_seconds | running_total | running_count | running_avg |
|---|---|---|---|---|---|
| 1 | 31000 | 74 | 74 | 1 | 74 |
| 2 | 31000 | 291 | 365 | 2 | 182.5 |
| 3 | 31000 | 520 | 885 | 3 | 295 |
| 4 | 31000 | 424 | 1756 | 5 | 351.2 |
| 5 | 31000 | 447 | 1756 | 5 | 351.2 |
| 6 | 31000 | 1422 | 3178 | 6 | 529.6666666666666 |
| 7 | 31000 | 348 | 3526 | 7 | 503.7142857142857 |

```
SELECT start_terminal,
       start_time,
       duration_seconds,
       ROW_NUMBER() OVER (PARTITION BY start_terminal
                          ORDER BY start_time)
                    AS row_number
  FROM tutorial.dc_bikeshare_q1_2012
 WHERE start_time < '2012-01-08'
```

| | start_terminal | start_time | duration_seconds | row_number |
|---|---|---|---|---|
| 1 | 31000 | 2012-01-01 15:32:... | 74 | 1 |
| 2 | 31000 | 2012-01-02 12:40:... | 291 | 2 |
| 3 | 31000 | 2012-01-02 19:15:... | 520 | 3 |
| 4 | 31000 | 2012-01-03 07:22:... | 424 | 4 |
| 5 | 31000 | 2012-01-03 07:22:... | 447 | 5 |
| 6 | 31000 | 2012-01-03 12:32:... | 1422 | 6 |
| 7 | 31000 | 2012-01-04 17:36:... | 348 | 7 |
| 8 | 31000 | 2012-01-05 15:13:... | 277 | 8 |